

# “Biblicality” of Early Medieval Canon Law through the Lens of Language Modeling

Friederike Voit<sup>1</sup>, Gleb Schmidt<sup>1</sup> , and Sven Meeder<sup>1</sup> 

<sup>1</sup> Rich Institute of Culture and History, Radboud University, Nijmegen, Netherlands

## Abstract

Using language modeling and the information-theoretic concept of perplexity, this study explores the influence of the Bible on early medieval canon law. We demonstrate that calculating perplexity of a corpus of canon law under a language model trained exclusively on Scripture can serve as a reliable proxy for accessing the overall linguistic similarity—stylistic, semantic, and syntactic—of canon law to the Bible. The paper presents the measured “biblicality” of various canon law texts, explores its chronological development, and delves into the linguistic and stylistic meaning of higher or lower “biblicality” by observing correlations between perplexity under the biblical language model and various linguistic features of canon law texts which can be extracted from rich morpho-styntactic annotation. We hypothesize that changes in the levels of “biblicality”, clearly observable across the chronological subdivisions of the corpus, suggest that the imitation of scriptural language may have been a deliberate strategy reflecting evolving views on the role and place of Scripture in legislation. Further research is needed to trace in more detail the connection between the authoritative status of canon law collections and the use of the Bible.

**Keywords:** Early Medieval Canon Law, Bible, Intertextuality, Language Modeling, Perplexity, Styliometry

## 1 Introduction

“... We condemn those who promote rules that go against Scripture and the ecclesiastical canons and introduce new precepts.<sup>1</sup>”

The ninth-century author of this quote points to a central issue for medieval *érudits* and, arguably, all of medieval society. Some texts carry authority, others do not. Distinguishing between these two categories was essential, especially in the context of normative texts such as canon law. Early medieval canon law is a rich, varied, and often unstructured body of texts which aims to communicate rules, norms, morals, and ideals for communities. Among authoritative sources of law, the Bible clearly took center stage, alongside church councils, papal letters, and the writings of a constantly redefined group of “Church Fathers”. Texts without authority included—equally obviously to the medieval mind—any novel rules by upstart authors.

When deciding whether a text had authority, be it legal, spiritual, political or social, medieval readers trusted clear attributions. Yet they were equally aware that such attributions were often flawed, references muddled, and quotations manipulated or even outright forged [17]. This raises the question: What other markers could have aided readers in assessing a text’s authority?

A comprehensive exploration of the medieval notion of authority and the factors that shaped it would be impossible within the scope of a single study, even one restricted to the genre of canon

Friederike Voit, Gleb Schmidt, and Sven Meeder. ““Biblicality” of Early Medieval Canon Law through the Lens of Language Modeling.” In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 192–211. <https://doi.org/10.63744/i0rY9iLox3mL>.

© 2025 by the authors. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).

<sup>1</sup> “Sed et hos condemnamus qui se extollunt adversus scripturas et ecclesiasticos canones et nova introducunt praecepta” (PL 84, col. 115). A ninth-century forger attributed this text to the Synod of Gangra (340), presenting it as its epilogue.

law such as ours. For this reason, in this paper we narrow our focus to a single problem: the textual influence that the Bible—the ultimate source of authority—exerted on the canon law corpus.

Importantly, by choosing this perspective, we do not want to imply that associating a canon law text with Scripture was necessarily a major—let alone the sole—means of establishing a text’s authority. Such a claim would be too bold and trivial at the same time. Rather, we aim at offering a global overview of the presence of the Bible within the canon law corpus, in order to determine whether this presence exhibited a discernible structure—chronological or topical—that could be interpreted historically. Such an interpretation, in turn, may—in the future—shed light on whether *biblicizing*, i.e., imbuing a text with a perceptible influence of the Bible—can indeed be considered as a conscious communicative strategy for constructing the authority of normative texts.

Our approach echoes recent scholarship on canonicity in literature, which focuses on the interpretation of linguistic features as determinants explaining why certain written works enter the literary canon thus achieving lasting cultural significance [3; 56; 58]. It also speaks—albeit more distantly—to M. M. Bakhtin’s concept of “dialogic literature”, in which the meaning of texts emerges from a continuous and reciprocal dialogue with other works [2].

Without presuming to answer the question of how exactly the Bible contributed to establishing the authority of normative texts, we wish simply to listen—following Bakhtin—to the dialogue between canon law and the Bible, in the hope of discerning whether this dialogue remained equally lively and intensive across the centuries.

Assessing “textual influence” of the Bible on canon law is not straightforward, as it may manifest at varied levels of granularity—from multi-word expressions, syntactic patterns, recurring allegories, commonplaces, or *topoi*— unsurprisingly, leaving the question open as to how it can be systematically studied.

A number of other factors also complicate the matter and must be taken into account when selecting a methodological approach. At the most general level, Scripture is the source of all that regulates Christian life, and the entire body of canon law is therefore, by definition, biblical—as the ninth-century quotation above illustrates. Yet the Bible contains few norms presented in a form that would have been directly applicable to early medieval life. Although there were attempts to derive such norms from Scripture—most notably by the Irish monks responsible for the *Collectio canonum Hibernensis* [17; 18; 59]—this approach was neither exclusive nor mainstream. More commonly, biblical norms entered canon law mediated through excerpts from patristic writings, theology, and political philosophy [38; 42]. Moreover, a text that “sounds” biblical in tone and register may reflect both the central role of the Bible in early medieval education and its status as the underlying linguistic framework of the literate Latin West [57].

To capture a variety of ways in which the influence of the Bible on canon law manifest, we propose using language modeling and a perplexity-based measure. As we hope to demonstrate, this approach makes it possible not only to quantify the degree of “biblicality”—the overall similarity to the Bible—of any given canon law text, but also to do so without relying on a pre-theoretical and thus bias-prone definition of what this “biblicality” entails. Instead, the interpretable evidence provided by our method can be analysed in its relevant context, contributing to a stronger understanding of the diverse law-making strategies employed by the compilers of early medieval canon law.

Building upon previous research, this paper introduces and critically evaluates a novel, corpus- and language-agnostic approach to intertextual association. Its main contributions can be summarized as follows:

1. Situating perplexity alongside other metrics, it identifies linguistic and semantic features that constitute “biblicality”;
2. Employing this approach, it identifies diachronical patterns in the influence of the Bible on

- early medieval canon law;
3. Using the case study of the *Collectio 250 capitulorum*, it explores the applicability of this approach *within* a canonical collection.

## 2 Related Work

The complexity of these questions explains why not all levels of textual influence have been considered equally in earlier scholarship. While the identification of quotations across corpora [35; 36], as well as allusion detection [32; 33] and thematic affinity analysis [10], have advanced, lower-level linguistic similarities—such as tone, register, and syntactic patterns—remain underexplored.

We believe that it is possible to fill this gap by taking inspiration from stylometry, historical linguistics, and literary studies, and adopting an approach based on generative language modeling. The idea of conducting text classification indirectly, through generative language modeling—the so-called “generative approach”—stems from classical machine learning. Rather than explicitly searching for features that distinguish between corpora (the “discriminative approach”), this approach models the language of each class in the corpus as a whole, as if the task were to generate new samples of that class. A well-trained generative model can then predict if unseen texts resemble those belonging to the class in question [43].

### 2.1 Perplexity-based Language Distances

In historical linguistics, generative language modeling has been used to “discriminate between similar languages or varieties and to measure language distance both synchronically and diachronically” [9]. Here, perplexity is a measurement of how well a model that has been trained on a corpus predicts a sample text [26]. Higher perplexity means that the model is more uncertain about its predictions, and that the sample text is less closely associated with the training corpus. Such a perplexity-based approach involves using a statistical language model—character- or word-*n*-gram-based models have been tested—which captures the language as a whole and is then used to compute a perplexity-based distance (see Section 3.2).

### 2.2 Perplexity as an Indication of Intra- and Intertextuality

Language modeling and perplexity-based distances have also been applied to research questions such as authorship analysis [1; 25], the relationship between languages [9; 20; 21; 22; 44], intertextuality and formulaic language [14; 28; 41], canonicity [61] and literary quality [29].

Recently, a number of studies used character *n*-gram statistical language models (SLMs) to estimate linguistic proximity between different books in the Homeric poems, exploring the internal structure of the poems with a focus on reoccurring formulae and possible later interpolations [14; 28; 40; 41]. In later studies [28; 41], the same approach was applied at the verse level. By computing a “cross score” for each verse (see Appendix E), Konstantinidou, Pavlopoulos, and Barker were able to identify verses (*ca.* 2–3% of the total number) which are linguistically surprising in their immediate context. Subsequently, they examined correlations between such “unexpectedness” and linguistic features of the verses, including the presence of named entities and *hapax legomena*.

## 3 Methodology

Leveraging these methods to explore the “biblicality” of early medieval canon law, we trained a language model exclusively on the Vulgate and calculated perplexity of canon law texts under this model. Lower perplexity under the biblical language model—indicating that these texts are less challenging for the model—was interpreted as an indicator of a stronger linguistic affinity with the

Bible, without *necessarily* implying philological or historical causation. Additionally, we trained a model on canon law to contextualize each canon law text within its own corpus.

### 3.1 Model

In principle, any generative model can be used to implement this approach. We opted for a compact character-level GPT-2, which we consider arguably the most powerful model available for training on limited data.

This choice was motivated by two factors. Firstly, given the size and heterogeneity of our corpus, we could not rely on direct estimation of joint  $n$ -gram probabilities Peng et al. or Fasoi, Pavlopoulos, and Konstantinidou did. Moreover, we sought to leverage more advanced deep learning techniques than those employed by Ge, Sun, and Smith and Bagnall, who successfully used feedforward and Recurrent Neural Networks (RNNs) respectively. The total size of our dataset, especially the biblical corpus, is too small for effective token-level modeling without overfitting, even when using a custom tokenizer with a limited-size vocabulary (see Section 4). Therefore, we considered character-level training a necessary trade-off, allowing us to leverage the robustness of GPT-2 in capturing regularities in texts, but at a more granular level. Although it is a considerable limitation, working at a character level did have advantages. Fine-grained patterns manifested in short spans of several words—at the sub-quotation level, so to speak—remain, as previously noted, among the most understudied dimensions of the Bible’s influence on canon law.

For implementation, training procedure, and hyperparameters, see Appendix B.1.

### 3.2 Perplexity-based Proximity Score

We employed neither the perplexity-based distance proposed by Campos, Gamallo, and Alegria [9] nor the positive cross-score introduced by Konstantinidou, Pavlopoulos, and Barker [28] (for formal definitions of these metrics, see Appendix E), but developed a formal scoring procedure—*Perplexity-based Proximity Score* (PPS)—to identify canon law texts which exhibit a stronger affinity to the Bible while being less typical of the canon law corpus.

PPS compares the  $z$ -scores of text perplexities under two models. Standard score normalization efficiently addresses the problem of scale differences (see Appendix D), allowing us to assess whether any given text is *relatively* more or less perplexing to a specific model, compared to the average perplexity of texts within the corresponding corpus and the same model. As a result, PPS helps to highlight canon law texts that are (1) less perplexing to the biblical model than canon law texts are on average, and (2) more perplexing to the canon law model than canon law texts are on average. Or, more formally:

$$PPS = z_{\text{test}}(\text{PP}_t) - z_{\text{biblical}}(\text{PP}_t),$$

where  $\text{PP}_t$  denotes the perplexity of text  $t$  under the specified model, and  $z(\text{PP})$  is the  $z$ -score of that perplexity. Only texts with a negative  $z$ -score under the biblical model are considered.

## 4 Data

Five bodies of data were used in this study. The Latin Vulgate Bible and a corpus of early medieval Latin canon law—currently under construction as part of “The Social Life of Early Medieval Normative Texts” (SOLEMNE) project—take center stage in our experiments. Additionally, for the validation procedure described in Section 5 below, we retrieved three supplementary corpora: (1) the *Latinitas antiqua* corpus from the *Corpus Corporum* [50], (2) Theodosius’ and Justinian’s

legislation extracted from the Latin Library [11], and (3) the developing Patristic Sermon Textual Archive (PaSTA),<sup>2</sup> which contains texts reflecting the homiletic tradition of the patristic era.

One of the ambitions of the ongoing SOLEMNE project is to provide access to the most comprehensive corpus of canon law from the fourth to the twelfth century and to equip it with the necessary heuristic and reference tools. At present, the corpus includes 415 texts and comprises 2,251,142 words. This encompasses: (1) published editions of councils, papal decretals, and canonical collections which are not yet available in digital form; (2) direct transcriptions of unpublished material; and (3) contributions from other comparable projects. It has greatly benefited from authorized reuse of data produced by several past and ongoing initiatives:

- the database of the Carolingian Canon Law (CCL) project [16];
- texts published by Michael Elliot’s Anglo-Saxon Canon Law project [13];
- XML-files shared by the Monumenta Germaniae Historica (MGH) [37].

For this study, the texts were assigned dates, provided with genre labels (council, decretal, penitential, chronological collection, systematic collection) and morpho-syntactic annotation was produced using Universal Dependencies EvaLatin 2024 [52]. For more corpus-related statistics, see Appendix A. For more details on preprocessing, see Section B.2 of Appendix B.

## 5 Validation

Applying concepts and methods from other fields, humanities scholars inevitably face numerous risks [12; 51]. As such, validation—preliminary checks to assess whether the applied scientific methods yield results that are both interpretable and relevant to the field—is essential. In order to assess whether perplexity measures correspond to features relevant to human readers, we applied a two-step validation procedure.

### 5.1 Cross-Corpus Validation

The first method involves testing a biblical language model against several corpora whose relationship to biblical language—or lack thereof—is well established for historical reasons. To this end, in addition to the models trained on the biblical and canon law datasets, we trained three further models based on the corpora described above in Section 4: Classical Latin, Patristic, and Roman Law. We then evaluated each model on every corpus, computing the perplexity of each corpus under each model. In this way, we assessed the relationships among these corpora through the lens of perplexity.

Due to the obvious risk of confirmation bias, such validation procedures often suffer from fundamental weaknesses and are rarely sufficient. However, we consider our case to be different. Each of the corpora represents a distinct genre or is thematically coherent: classical Roman prose, Roman law, and patristic preaching. Clear distinctions between them—as well as the intuitiveness with which any reader of Latin would describe their mutual relationships and degrees of similarity—are strong enough to make this a useful first step of validation.

### 5.2 Classification-based Validation

Our second validation procedure is based on classification. The assumption is that canonical texts which yield lower perplexity scores under the biblical model—and thus appear more biblical compared to other canon law texts—will also be more difficult to classify as part of the canon law

---

<sup>2</sup> An ongoing effort following the ERC Starting Grant “Patristic Sermons in the Middle Ages” (2019–2023, PI Dr. S. Boodts) [5; 7; 31].

corpus. To verify this, we trained three binary classifiers on canon law texts whose perplexity  $z$ -score under the biblical model is negative. We then identified the misclassified texts—i.e., canon law texts that the model incorrectly classified as biblical—and computed the correlation between the confidence of the classifiers’ incorrect prediction and the absolute perplexity of these texts under the biblical model. For a detailed overview of the procedure, see Appendix C.

### 5.3 Validation Results

Model/Corpus	Bible	Canon law	Roman law	Classical Latin	Patristic preaching
<b>Bible</b>	−1.79	0.29	0.84	0.95	−0.29
<b>Canon law</b>	0.14	−1.43	0.15	1.63	−0.50
<b>Roman law</b>	0.95	−0.52	−1.70	0.87	0.41
<b>Classical Latin</b>	1.92	−0.77	0.05	−0.66	−0.53
<b>Patristic preaching</b>	−0.63	−0.71	0.88	1.50	−1.03

**Table 1:** Cross-corpus validation. Standard-score-normalized perplexities of five GPT-2 models (rows), each evaluated on its source corpus (diagonal cells) and on four remaining corpora (off-diagonal cells). See Appendix D for more details, raw perplexity values (Table 5), and guidance on interpretation.

Table 1 and 5 (in Appendix D) present the perplexities of the five models, each tested on its source dataset and the four remaining corpora. Several observations indicate that this first validation check was confidently passed.

All five models performed best when predicting texts from their respective source datasets. Other patterns also align well with scholarly consensus. For example, the Roman law model performs well on canon law, which is the only other corpus that is thematically related to, and historically connected with, the texts used to train this model [4; 24; 30]. The patristic preaching model also performs well on the Bible. This likely reflects the inherently biblical focus of preaching in the patristic era [6]. Quoting scripture was not merely a rhetorical device but also integral to the liturgical rite [39]. As a result, the corpus that was used to train this model contains numerous direct references to the Bible.

One result stands out: the Classical Latin model. However, its surprisingly strong performance on canon law and patristic preaching may be explained by the somewhat unusual decision of the *Corpus Corporum*’s curators to include within this *Latinitas Antiqua* corpus authors from an excessively broad period even including Church Fathers such as Jerome, Bede, and Augustine.

Classifier	Precision	Recall	F1	Accuracy	Spearman’s $\rho$	Pearson’s $\rho$	Kendall’s $\tau$
RandomForestClassifier	0.84	0.87	0.85	0.85	Spearman’s $\rho$	−0.20	
SVC	0.83	0.73	0.77	0.79	Pearson’s $\rho$	−0.21	
KNeighborsClassifier	0.88	0.76	0.82	0.83	Kendall’s $\tau$	−0.14	

**Table 2:** Classification-based Validation

Table 2 summarizes the results of classification-based validation. It shows the performance of binary classifiers trained to recognize canon law texts (positive class), along with three correlation coefficients computed across all classifiers,<sup>3</sup> capturing the relationship between perplexity under the biblical model and the confidence with which samples were misclassified. Notably, all three

<sup>3</sup> In all three cases, the p-value is below 0.0001.

		pearson	0.69
	kendalltau	0.81	0.61
average_word_length	0.51	0.38	0.54
average_tree_depth	0.38	0.52	0.53
average_sentence_length	0.38	0.51	0.51
nominal_style	0.36	0.63	0.48
nonfinite_verb_ratio	0.35	0.59	0.49
passive_voice_ratio	0.35	0.52	0.44
ADV	0.31	0.24	0.42
Mood=Sub	0.22	0.21	0.35
NOUN	0.22	0.21	0.31
ablativus_absolutus	0.22	0.21	0.28
VERB	0.11	0.10	0.26
relative_clause	0.09	0.10	0.20
SConJ	0.09	0.09	0.25
ADP	-0.12	-0.12	-0.13
CCONJ	-0.20	-0.22	-0.20
PRON	-0.22	-0.22	-0.34
PART	-0.26	-0.26	-0.30
quotation_marker_ratio	-0.30	-0.30	-0.47
Mood=Ind	-0.33	-0.34	-0.58
richness_Yule_k	-0.39	-0.39	-0.61
Mood=Imp	-0.65	-0.65	-0.52

**Figure 1:** Linguistic meaning of “biblicality”.

classifiers consistently yield a negative correlation—canon law texts that are misclassified as biblical (i.e., assigned a high negative-class probability) tend to have lower perplexity under the biblical model. The fact that all three values are weak to moderate should not surprise us: similarity to the Bible is only one possible reason for misclassification. Overall, we can be reasonably confident that perplexity under the biblical model correlates to canonical “biblicality”.

## 6 Experiments and Results

Following our validation procedures, we employed the trained biblical and canon law models for examination of our canon law corpus. The study is organized around three experiments.

### 6.1 Linguistic and Semantic Meaning of “Biblicality”

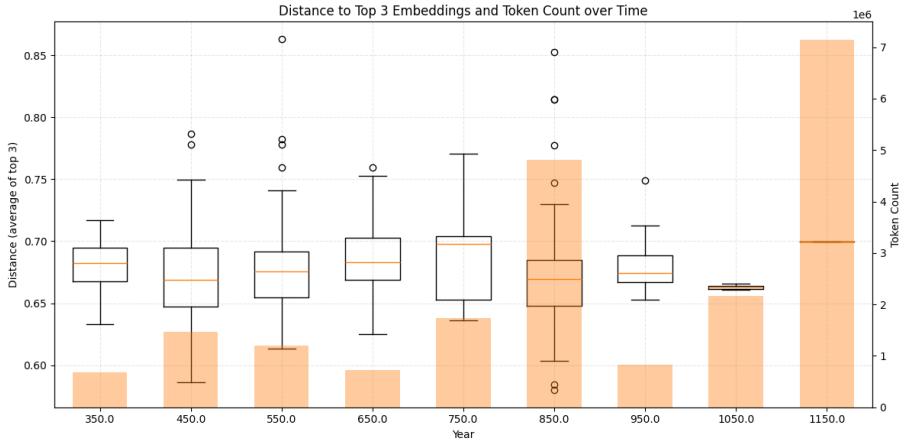
As previously discussed, a significant advantage of employing perplexity is that it avoids a pre-theoretical interpretation of “biblicality”. Yet this method also raises the question: what makes a text sound biblical?

Our validation procedures suggest that lower perplexity under the biblical model indeed reflects textual similarity to the Bible. In this first experiment, we attempt to identify specific linguistic and semantic features which constitute this similarity. The experiment has two parts. Firstly, we explore how perplexity correlates with various linguistic features.

We defined a number of linguistic features that can be easily extracted from an automatic morpho-syntactic annotation of the text. At the core of this feature set are the distributions of parts of speech, along with several traditional stylometric features. However, following the approach described by Ph. [48], we supplemented this feature set with several composite features that capture the density of information-rich and formal writing,<sup>4</sup> as well as *quotation\_marker\_ratio*, which reflects the prevalence of syntactic patterns typically associated with direct speech or quotation. Having extracted the features for each text, we computed the correlation coefficients between each feature and the perplexity under the biblical model.

As shown in Figure 1, “perplexity”, and its inverse “biblicality”, aligns with specific linguistic patterns, primarily along the axis of syntactic and lexical complexity. Texts with longer, more

<sup>4</sup> nominal\_style\_ratio, average\_tree\_depth, nonfinite\_verb\_ratio.



**Figure 2:** Perplexity and Median Distance of Closest Three Embeddings (100-year Spans)

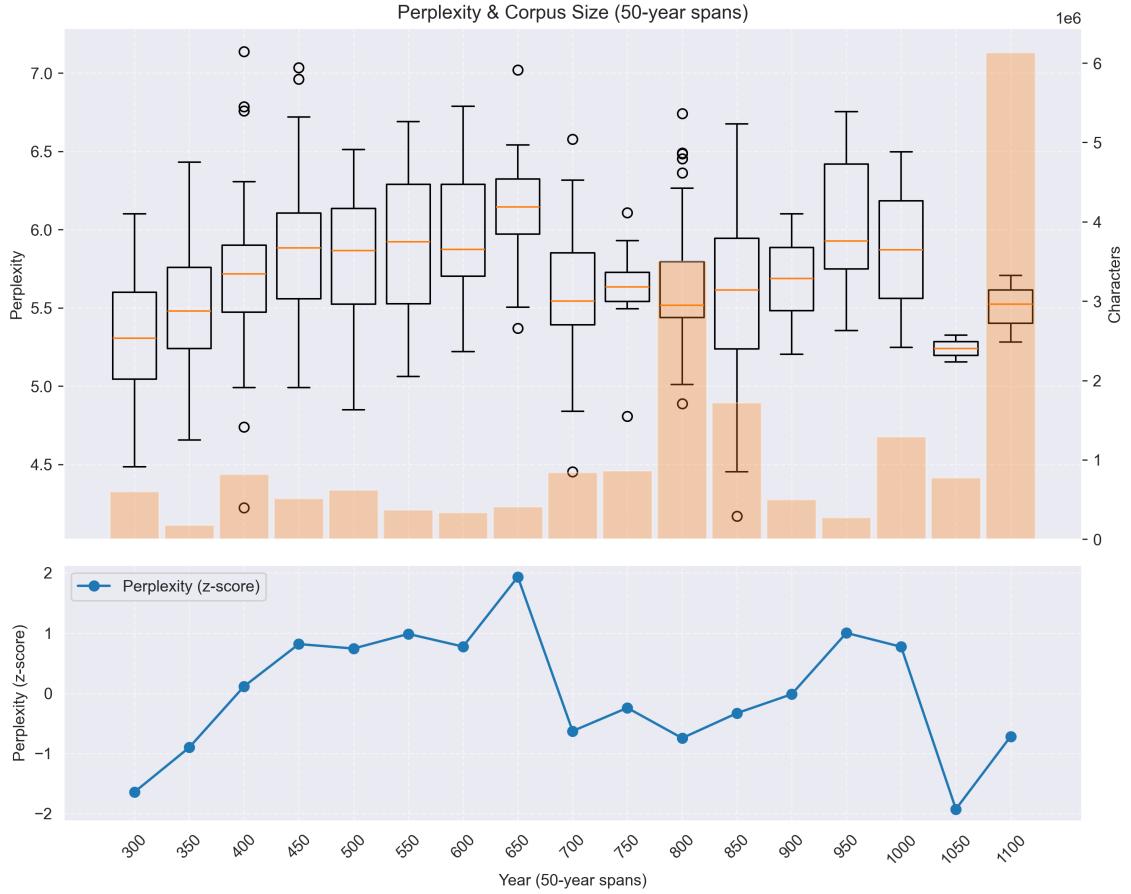
complex sentences, higher `nominal_style_ratio`, frequent passive constructions, and greater lexical diversity (measured through `richness_Yule_k`, where a lower index corresponds to a richer vocabulary) tend to be more perplexing under the biblical model. This will prove especially relevant for our assessment of papal decretals, a sub-category of canon law which is often syntactically complex due to its epistolary form [16; 27].

Conversely, texts with more coordinating structures and frequent use of the indicative and imperative moods are less perplexing under the model. The feature `quotation_marker_ratio` is also revealing. High frequency of quotation markers corresponds to lower perplexity under the biblical model, possibly reflecting the extent to which texts directly quote Scripture. Since this feature captures constructions marking reported speech or communication, it may encompass both acts of speech within excerpted passages or the attribution of quotations to an authority such as the Bible. This result is also intriguing given that attributions through formulaic speech markers are a common convention in canon law texts, especially systematic canonical collections [19].

Beyond correlating perplexity with different linguistic features, the second part of this experiment examines how perplexity relates to the semantic aspects of our canon law corpus. We employ the `bowphs/LaBerta` model [47] to obtain semantic representations for every Bible verse and every fragment of canon law text. Subsequently, for each vector representing a canon law fragment, we compute the median distance to its three closest Bible verses. This yields an alternative, primarily semantic measure of “biblicality”. Figure 2 presents a diachronic comparison of the  $z$ -scores for this embedding-based measure and those of perplexity.

The embedding-based measure has a broadly similar trend to perplexity. Both measures highlight three periods of lower “biblicality”, around 600-700, 900 (although this is less apparent in the embedding-based trend line), and after 1100. Additionally, both measures indicate higher “biblicality” at similar points, such as around 800. This suggests that a significant part of what produces high or low perplexity under our biblical model is the semantic closeness of texts to the Bible. This may encompass direct quotations (which are semantically near-identical to the Bible), but also captures biblical allusions and statements whose content merely resembles the Bible. This provides a useful starting point for our analysis of “biblicality” in canon law texts such as penitentials, which contain fewer direct quotations of the Bible.

By investigating how perplexity corresponds to linguistic and semantic features, we thus reach initial answers about how canon law texts assert their similarity to the Bible. In particular, our comparison of perplexity with an embedding-based measure hints at historically significant trends in the diachronic relationship of canon law texts to the Bible. This provides the foundation for our analysis of how the “biblicality” of canon law changes throughout the early Middle Ages.



**Figure 3:** “Biblicality” of Canon Law in the Fourth-Twelfth Centuries

## 6.2 Diachronic “Biblicality” of Canon Law, ca. 300-1150

The second experiment examines the diachronic trajectory of the Bible in our canon law corpus. We test a long-standing hypothesis, originally proposed by Jean Gaudemet [23], that the Bible played little role in canon law texts from the sixth to mid-ninth centuries and that it was only during the Carolingian era that it became, to borrow Contreni’s expression, “a prominent part of the landscape informing contemporary thought” [54] in canon law.

A secondary goal of this experiment is to highlight an underexplored feature of early medieval legal culture. Abigail Firey recently discussed the essential peculiarity of canon law in the early Middle Ages. She argued that it “accommodated pluralism [...] Such untidiness does not seem to have unduly hindered those seeking to learn and apply canon law, or to have damaged confidence in its value” [15, p. 13]. Our investigation sheds light on this legal pluralism by foregrounding the coexistence of many discourse universes and strategies of asserting “biblicality”.

Figure 3 shows the median perplexity of canon law texts under the biblical model, distributed over chronological bins of 50 years. A clear increasing trend from *ca.* 300 to 650 reflects gradually increasing perplexity and decreasing “biblicality” over this period. This aligns with Gaudemet’s hypothesis, and indeed, it is historically intuitive. In the fourth century, Jerome translated the Hebrew and Greek Bible and revised earlier Latin versions, producing the Vulgate Bible. As such, the language of the Vulgate resembles that of other texts from this period [8; 55]. We see here not only the reflection of “biblicality” in early canonical works, but the reflection of underlying cultural and linguistic frameworks—including canon law—in Jerome’s education and work.

A period of significantly lower perplexity begins *ca.* 700. Further supporting Gaudemet’s

theory, the canon law texts produced at this time display a growing affinity with the Bible. This coincides with the composition of the *Collectio canonum Hibernensis*, which saw an expansion of the source materials used in canon law to include the Bible itself as well as highly biblical patristic writings [19; 46]. This trend of increasing “biblicality” continues throughout the eighth and ninth centuries. This period is equally significant in the history of canon law, as the forgers in the Pseudo-Isidorian workshop crafted their false decretals *ca.* 850. These texts contain innumerable biblical allusions[16] and—in their efforts to appear genuine—may have deliberately echoed the language of earlier popes, and thus also Jerome and the Vulgate. As such, the trend indicates that these influential canonical collections had a significant impact on the “biblicality” of canon law.

Finally, following a temporary increase over the tenth century, perplexity drops abruptly *ca.* 1050 and remains low into the twelfth century. This dramatic increase in “biblicality” coincides with the composition of a highly influential canonical collection, Gratian’s *Concordia discordantium canonum*. This collection incorporates a significant amount of biblical and patristic material. Indeed, Gratian’s work is known to have shifted the direction of canon law [45; 53]. Due to its considerable length, this work also constitutes the vast majority of our data for this period, which may also have amplified its influence on the trend. Nevertheless, this aspect of our analysis extends Gaudemet’s argument by hinting at the continuities and shifts in the “biblicality” of canon law going into the high and late Middle Ages.

This complex diachronic trajectory also speaks to the legal pluralism of the period [15]. Far from a linear relationship, it appears that historical events and processes, cultural and linguistic shifts, and developments within canon law itself all impacted the “biblicality” of the corpus at various times. Moreover, several chronological bins—especially *ca.* 400 and 700-800—contain a number of outliers with very high or low perplexity under the biblical model. Beyond diachronic developments, this highlights that canon law was always heterogeneous and pluralistic with regard to its “biblicality”.

Pluralism does not mean chaos. Some of these outliers can be explained by other factors which intersect with diachronic developments. Analyzing the 50 texts with the highest PPS, it is notable that 18 are decretals. This is a striking proportion, given that the same list includes only six chronological or systematic collections and a single penitential. Moreover, several of these decretals originate from the fifth and sixth centuries, when “biblicality” was generally low. This can plausibly be explained by the fact that decretals, as previously mentioned, tend to contain complex language and a high number of biblical references. As such, these results suggest that genre was a further factor influencing the prominence of the Bible in early medieval canon law.

More tentatively, geographic patterns can also be identified in the collections with remarkably high “biblicality”. In the ninth century, a period where the Bible was already highly present in canon law, no less than three Councils of Quierzy and two Councils of Aachen appear among the exceptionally “biblical” outliers. These locations are in close geographical proximity, suggesting that local cultural frameworks or perhaps even individual bishops or rulers may have played a role in the prominence of the Bible in some canonical collections. In this way, a diachronic analysis of “biblicality” provides a window into the legal pluralism of the early Middle Ages, foregrounding diversity while hinting at the underlying cultural, linguistic, and political structures that guided the development of canon law.

### 6.3 *Collectio 250 capitulorum*

Our third experiment analyses how diverse strategies of asserting “biblicality” converge in an individual canonical collection. The *Collectio 250 capitulorum* provides an intriguing example. Composed during the peak of highly “biblical” canon law in the ninth-century, this systematically arranged collection consists almost exclusively of excerpts from another highly biblical collection, the *Collectio canonum Hibernensis* [46; 59]. Indeed, its compilers emphasised the significance of

the Bible as a canonical authority. The opening canon, an excerpt from a papal decretal by Innocent I, directs readers towards:

[...] the twenty-five books of the Old Testament, and the four Gospels. When an answer should not present itself in all the writings of the apostles, turn to the divine Greek texts, i.e. the oecumenical councils. If you do not find the answer in them, reach for the catholic histories of the catholic church, [...] the canons of the apostolic see [...] the examples of saints [...] the elders of the province [...] For the true surely, the Lord, said: Should two of you or three come together upon the earth in my name, concerning anything whatsoever they shall ask, it shall be done for them [Matthew 18:19].<sup>5</sup>

This positions the Bible as the principal source of legal authority, relegating patristic writings, the *Canones Apostolorum*, and conciliar rulings to a supplementary role [34]. Concluding with a quotation from the Gospel of Matthew, it also exemplifies how biblical rulings could be used to interpret and reaffirm other authoritative sources. According to the compilers of the *Collectio 250*, canonical authority begins with, and inevitably circles back to, the Bible.

This raises the question of how the compilers' perspective on "biblicality" shaped the collection. In order to analyse this, we encode a transcription of one manuscript, München, Bayerische Staatsbibliothek, Clm 4592, according to the SOLEMNE XML-TEI standard and produce perplexity scores for each atomic unit (smallest citable chunk of text). The results highlight the collection's remarkable "biblicality". Of its 508 atomic units, 130 had a positive PPS, indicating that they were on average more difficult for the canon law model to predict and easier for the biblical model. Moreover, in 76 cases, the perplexity  $z$ -score under the biblical model was below -0.25 while the perplexity  $z$ -score under the canon law model was above 0.25. As such, the *Collectio 250* not only demonstrates high "biblicality" in a significant number of atomic units, but also to a statistically significant extent. This aligns with broader interpretations of how legal scholars in the Carolingian consciously evoked the Bible in their work [17; 19]. The compilers of the *Collectio 250* succeeded in their stated intention to compose "biblical" canon law.

Notably, the compilers also intensified their efforts in certain parts of the collection. Almost half of the 25 atomic units with the highest PPS concern judgements (six) or inheritance (six), while others discuss usury and stolen goods. This distribution is atypical for the collection, suggesting that the compilers considered "biblicality" more appropriate – or necessary – in financial and jurisprudential matters. Such insights expand existing discussions on the use of the Bible alongside other sources of canon law and perceived hierarchies of authority [17; 34]. It suggests that particular sources were considered more authoritative on certain topics, with compilers' approaches to the incorporation of biblical material shifting depending on subject matter. Moreover, highly "biblical" atomic units which are not attributed to the Bible itself are almost exclusively identified with the church fathers Jerome and Origen. These results are also unrepresentative of the collection—after attributions to Jerome (113), the most frequently cited patristic source is Augustine (68), with Origen only appearing on 27 occasions. This adds nuance to Jean Werckmeister's argument that "the Fathers were cited in their capacity as commentators of the Bible" [60]. Certain patristic authorities, such as Origen, were indeed disproportionately cited for their biblical material, while others like Augustine appear to have been cited for other reasons. This demonstrates that compilers took nuanced approaches to "biblicality" depending on the content and context of individual canons.

<sup>5</sup> "[...] XXV librorum veteris testamenti, IIII euangeliorum. Cum scriptis totis apostolorum non appareat, ad diuina recurrito scripta grece. Si nec in illis, ad catholicae ecclesiae historias catholicas [...] canones apostolicæ sedis [...] sanctorum exempla [...] seniores prouinciae [...] Uerus enim re promissor Dominus ait: Si duo ex uobis uel tres conueniant super terram in nomine meo de omni re quaecumque petierint fiet illis". For translation see Meeder [34].

Finally, analysis of the *Collectio 250* provides a glimpse into how compilers achieved this complex engagement with the Bible. The chapters containing the 25 atomic units with the highest PPS encompass 37 identifiable biblical references and 30 other statements.<sup>6</sup> Many of the identifiable references are not quotations, but abbreviations of biblical passages. This underscores that direct quotation was not the only—and perhaps not even the most effective—strategy to assert affinity with the Bible.

Indeed, the chapter containing the atomic unit with the single highest PPS reveals compilers' layered engagement with biblical syntax, *exempla*, and quotations. Eight of its eleven statements are correctly attributed references to the Bible. Two further statements are (falsely) attributed to the apostle Paul, guiding readers to read them as biblical excerpts. Notably, all statements are paraphrased in clear and repetitive language, retaining the names of biblical characters but employing an otherwise limited vocabulary (e.g. frequent repetition of *iurare/iuramentum*, “to judge/judgement”) and *testari/testimonium*, “to testify/testimony”). This syntax and vocabulary further simplifies the already uncomplicated language of the Bible, effectively creating ultrabiblical Bible quotations. In this way, perplexity-based analysis also facilitates insights into the layered strategies by which compilers foregrounded the Bible in their work. Even within a single collection, “biblicality” had a range of meanings and was achieved in various ways.

## 7 Conclusion

Overall, perplexity-based analysis has proved useful in modeling the influence of the Bible on early medieval canon law. This approach facilitated analysis of the linguistic and semantic features that signal biblical affinity, without imposing preconceived assumptions. On a diachronic scale and at the granular level of individual texts, our PPS metric of similarity enabled us to analyse the “biblicality” of the corpus of canon law.

Furthermore, the results facilitated intriguing historical insights. Analysis of the chronological trajectory of the Bible in canon law substantiated a Gaudemet's hypothesis regarding the greater “biblicality” of canon law in the fourth and ninth centuries. Moreover, these threads of biblical influence provided a window into the legal pluralism of the Carolingian era and the early Middle Ages more generally. Our experiments also revealed different strategies of asserting “biblicality” in canon law. We identified a range of linguistic features which contribute to canon law texts sounding biblical, and analysed how these features were applied and combined in one highly biblical collection. This ultimately highlighted medieval legal scholars' diverse understandings of both “biblicality” and the norms of canon law.

Although this study revealed trends in the use of “biblicality” as a linguistic strategy, it remains unclear how—and whether—this “biblicality” was successful in communicating authoritativeness. Unlike researchers of (early) modern “canonicity” and cultural canons, those studying the early Middle Ages cannot draw on publishing data, sales records, and historical catalogs to test hypotheses about popularity and authority. Since our corpus of canon law depends largely on accidents of survival, care must be taken not to conflate the preservation of texts with their perceived authoritativeness in the early Middle Ages. Future historical research is needed to address this problem.

The study has also demonstrated the need for further corpus expansion and experimentation. Most notably, the Bible may not appear monolithically in canon law. The investigation of linguistic and semantic meaning revealed numerous elements that comprise “biblicality”, while analysis of the *Collectio 250 capitulorum* highlighted compilers' nuanced and selective use of biblical material. It would be valuable to train models on individual biblical books, investigating the comparative form and weight of their influence. Additionally, diachronic analysis revealed several canon

<sup>6</sup> Identification of biblical reference is based on Roy Flechner's edition of the *Hibernensis* [18].

law texts—notably decretals—with an unusually low perplexity  $z$ -score under the biblical model. While a key limitation of our corpus was the near-absence of some types of canon law, such as penitentials, an expansion of the dataset and an exploration of biblical influence on various types of canon law also seems promising.

**Another tip.** In some cases, it may be helpful to use `paragraph` to title individual paragraphs. For example, if a section describes features for a classifier, you can optionally title each paragraph with the name of each feature.

## Acknowledgements

An earlier version of this paper was presented at the conference for “Formulaic Language in Historical Linguistics: data, methods, tools, and theory” in Helsinki (2–3 June 2025). We thank the organisers for their invitation and the attendees for their thoughtful questions.

We extend our gratitude to the anonymous reviewers for their attentive reading and insightful suggestions, which contributed significantly to the final version of this study.

## Funding

Funded by the European Union. Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. This work is supported by the SOLEMNE project (ERC Cons. grant 101087979).

## References

- [1] Bagnall, Douglas. “Authorship clustering using multi-headed recurrent neural networks”. In: (2016), n.p. DOI: 10.48550/ARXIV.1608.04485. (Visited on 12/05/2023).
- [2] Bakhtin, Mikhail Mikhailovich. *The dialogic imagination: Four essays*, ed. by Michael Holquist. Trans. by Emerson Caryl and Michael Holquist. Austin, TX: University of Texas Press, 1981.
- [3] Barré, Jean, Camps, Jean-Baptiste, and Poibeau, Thierry. “Operationalizing Canonicity”. In: *Journal of Cultural Analytics* 8, no. 1 (2023), p. 88113.
- [4] Biondo, Biondi. *Diritto Romano cristiano*. Milano: Giuffrè, 1952.
- [5] Boodts, Shari. “Navigating the Vast Tradition of St. Augustine’s Sermons”. In: *Augustiniana* 69, no. 1 (2019), pp. 83–115. DOI: 10.2143/AUG.69.1.3286703.
- [6] Boodts, Shari and Schmidt, Gleb. “Sermon/Homiletics”. 2022.
- [7] Boodts, Shari, Schmidt, Gleb, Macchioro, Riccardo, Denis, Iris, Rempt, Menna, Komen, Erwin, and Hermsen, Thijs. “PASSIM Research Tool”. In: (2024). Publisher: Radboud Humanities Computing Lab. URL: <https://passim.rich.ru.nl>.
- [8] Brown, Dennis. “Jerome and the Vulgate”. eng. In: *A history of biblical interpretation. 1: The ancient period*, ed. by Alan J. Hauser. Num Pages: 536. Grand Rapids: Eerdmans, 2003, pp. 355–379. ISBN: 978-0-8028-4273-2.
- [9] Campos, José Ramon Pichel, Gamallo, Pablo, and Alegria, Inaki. “Measuring language distance among historical varieties using perplexity. Application to European Portuguese.” In: *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*. 2018, pp. 145–155.

- [10] Camps, Jean-Baptiste, Baumard, Nicolas, Langlais, Pierre-Carl, Morin, Olivier, Clérice, Thibault, and Norindr, Jade. "Make love or war? monitoring the thematic evolution of medieval french narratives". In: *Computational Humanities Research (CHR 2023)*. Paris, 2023.
- [11] Carey, William L. "The Latin Library". 2025. URL: <https://www.thelatinlibrary.com> (visited on 10/28/2025).
- [12] Da, Nan Z. "The Computational Case against Computational Literary Studies". en. In: *Critical Inquiry* 45, no. 3 (Mar. 2019), pp. 601–639. ISSN: 0093-1896, 1539-7858. DOI: 10.1086/702594.
- [13] Elliot, Michael. "Anglo-Saxon Canon Law". URL: <http://individual.utoronto.ca/michaellelliot/> (visited on 07/18/2025).
- [14] Fasoi, Maria, Pavlopoulos, John, and Konstantinidou, Maria. "Computational Authorship Analysis of Homeric Language". en. In: *Digital Humanities Workshop*. Kyiv Ukraine: ACM, Dec. 2021, pp. 78–88. ISBN: 978-1-4503-8736-1. DOI: 10.1145/3526242.3526256.
- [15] Firey, Abigail. "Between Chaos and Codification: Consensus and the Content of Carolingian Canon Law". In: *Standardization in the Middle Ages*, ed. by Line Cecilie Engh and Kristin B. Aavitsland. De Gruyter, Nov. 2024, pp. 13–42. ISBN: 978-3-11-098712-6. DOI: 10.1515/9783110987126-002.
- [16] Firey, Abigail. "Lawyers and Wisdom: The Use of the Bible in the Pseudo-Isidorian Forged Decretals". en. In: *Medieval Church Studies*, ed. by Celia Chazelle and Burton Van Name Edwards. Vol. 3. Turnhout, Belgium: Brepols Publishers, Jan. 2003, pp. 189–214. DOI: 10.1484/M.MCS-EB.3.3564. (Visited on 05/24/2025).
- [17] Flechner, Roy. *Making Laws for a Christian Society: The Hibernensis and the Beginnings of Church Law in Ireland and Britain*. en. 1st ed. Milton Park, Abingdon, Oxon ; New York, NY : Routledge, 2021. | Series: Studies in early medieval Britain and Ireland: Routledge, 2021. DOI: 10.4324/9781351267243. (Visited on 07/13/2022).
- [18] Flechner, Roy. *The Hibernensis: A Study and Edition*. Series Title: Studies in Medieval and Early Modern Canon Law. Washington DC: Catholic University of America Press, 2019.
- [19] Flechner, Roy. "The problem of originality in early medieval canon law: legislating by means of contradictions in the Collectio Hibernensis". In: *Viator* 43, no. 2 (2012), pp. 29–47.
- [20] Gamallo, Pablo, Alegria, Inaki, Campos, José Ramom Pichel, and Agirrezabal, Manex. "Comparing two basic methods for discriminating between similar languages and varieties". In: *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*. 2016, pp. 170–177.
- [21] Gamallo, Pablo, Pichel, José Ramom, and Alegria, Iñaki. "From language identification to language distance". In: *Physica A: Statistical Mechanics and its Applications* 484 (2017), pp. 152–162.
- [22] Gamallo, Pablo, Pichel, José Ramom, and Alegria, Iñaki. "Measuring Language Distance of Isolated European Languages". en. In: *Information* 11, no. 4 (Mar. 2020), p. 181. ISSN: 2078-2489. DOI: 10.3390/info11040181.
- [23] Gaudemet, Jean, Riché, Pierre, and Lobrichon, Guy. *La Bible dans les collections canoniques*. Paris: Beauchesne, 1984.
- [24] Gaudemet, Jean-Philippe. "Le droit romain dans la pratique et chez les docteurs aux XIe et XIIe siècles". fre. In: (1965). Publisher: Persée - Portail des revues scientifiques en SHS. DOI: 10.3406/ccmed.1965.1350.

- [25] Ge, Zhenhao, Sun, Yufang, and Smith, Mark. “Authorship attribution using a neural network language model”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 30. Issue: 1. 2016, pp. 4212–13.
- [26] Goldberg, Yoav. *Neural network methods for natural language processing*. Synthesis Lectures on Human Language Technologies (SLHLT). Berlin: Springer Nature, 2017.
- [27] Jasper, Detlev. *Papal letters in the Early Middle Ages*. eng. History of medieval canon law. Washington, D.C: Catholic University of America Press, 2001. ISBN: 978-0-8132-0919-7.
- [28] Konstantinidou, Maria, Pavlopoulos, John, and Barker, Elton. “Exploring intertextuality across the Homeric poems through language models”. In: *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*. 2024, pp. 260–268.
- [29] Kurzynski, Maciej. “Perplexity Games: Maoism vs. Literature through the Lens of Cognitive Stylometry”. en. In: *Journal of Data Mining & Digital Humanities NLP4DH* (2024), p. 13131. ISSN: 2416-5999. DOI: 10.46298/jdmdh.13131.
- [30] Maassen, Friedrich. *Geschichte der Quellen und der Literatur des canonischen Rechts im Abendlande bis zum Ausgang des Mittelalters*. Graz, 1870.
- [31] Macchioro, Riccardo et al. “The PASSIM Project (Patristic Sermons in the Middle Ages): Towards a Virtual Research Environment for the Study of Patristic Sermon Collections”. In: *CLASSICS@* (2021).
- [32] Manjavacas, Enrique. *Computational Approaches to Intertextuality*. PhD thesis. Antwerp University, 2021.
- [33] Manjavacas, Enrique, Long, Brian, and Kestemont, Mike. “On the Feasibility of Automated Detection of Allusive Text Reuse”. In: (May 2019). DOI: arXiv:1905.02973.
- [34] Meeder, Sven. “Negotiating Biblical Authority in the Collectio 250 capitulorum and the Collectio 400 capitulorum”. In: *The Politics of Interpretation: The Bible and the Formation of Legal Authority in the Early Middle Ages*, ed. by Gerda Heydemann and Rosamond McKitterick. Quellen und Forschungen zum Recht im Mittelalter 17. Ostfildern: Jan Thorbecke Verlag, 2025, pp. 81–100. ISBN: 978-3-7795-6097-9.
- [35] Mellerin, Laurence. “Biblindex”. In: *Bulletin de l'Association des Amis de «Sources Chrétiennes»* , no. 110 (2019), pp. 35–36.
- [36] Mellerin, Laurence. “Biblindex: An Index of Biblical Quotations in Early Christian Literature”. In: *The Harp. A Review of Syriac and Oriental Ecumenical Studies* 37 (2021), pp. 449–510.
- [37] “Monumenta Germaniae Historica”. URL: <https://www.mgh.de/en/digital-mgh/dmgh/linking-and-citing-dmgh> (visited on 05/21/2025).
- [38] Moore, Michael Edward. “The Ancient Fathers: Christian Antiquity, Patristics and Frankish Canon Law.” In: *Millennium Yearbook/Millennium-Jahrbuch* 7, no. 1 (2010), pp. 293–342.
- [39] Olivar, Alexandre. *La predicación cristiana antigua*. spa. Biblioteca Herder 189 Sección de teología y filosofía. Barcelona: Ed. Herder, 1991. ISBN: 978-84-254-1715-3.
- [40] Pavlopoulos, John and Konstantinidou, Maria. “Computational Authorship Analysis of the Homeric Poems”. en. In: *International Journal of Digital Humanities* (July 2022), n.p. DOI: 10.1007/s42803-022-00046-7.
- [41] Pavlopoulos, John, Sandell, Ryan, Konstantinidou, Maria, and Bozzone, Chiara. “HoLM: Analyzing the Linguistic Unexpectedness in Homeric Poetry”. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 2024, pp. 8166–8172.

- [42] Pelland, Gilles. “Incidence de l’exégèse sur l’évolution du droit canonique durant la première partie du Moyen Age”. In: *Periodica de re canonica* 82, no. 1 (1993), pp. 9–25.
- [43] Peng, Fuchun, Schuurmans, Dale, Wang, Shaojun, and Keselj, Vlado. “Language independent authorship attribution using character level language models”. en. In: *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - EACL ’03*. Vol. 1. Budapest, Hungary: Association for Computational Linguistics, 2003, pp. 267–74. DOI: 10.3115/1067807.1067843.
- [44] Pichel Campos, Jose Ramon, Gamallo, Pablo, and Alegria, Iñaki. “Measuring language distance among historical varieties using perplexity. Application to European Portuguese.” In: *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, ed. by Marcos Zampieri, Preslav Nakov, Nikola Ljubešić, Jörg Tiedemann, Shervin Malmasi, and Ahmed Ali. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 145–155.
- [45] Rennie, Kriston R. *Medieval canon law*. Past imperfect. Leeds, UK: ARC Humanities Press, 2018. ISBN: 978-1-942401-68-1.
- [46] Reynolds, Roger E. “Unity and Diversity in Carolingian canon law collections: the case of the Collectio Hibernensis and its derivatives”. In: *Law and Liturgy in the Latin Church, 5th-12th Centuries*. 1983, pp. 99–135.
- [47] Riemenschneider, Frederick and Frank, Anette. “Exploring Large Language Models for Classical Philology”. en. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, 2023, pp. 15181–15199. DOI: 10.18653/v1/2023.ac1-long.846. (Visited on 02/10/2024).
- [48] Roelli, Philipp. *Latin as the Language of Science and Learning*. Lingua Academica Beiträge zur Erforschung historischer Gelehrten und Wissenschaftssprachen 7. Berlin/Boston: De Gruyter, Nov. 2021. ISBN: 978-3-11-074583-2.
- [49] Roelli, Philipp. “On the Usability of Available Digital Tools for Reconstructive TextualEditing”. en. In: *Journal of Data Mining & Digital Humanities On the Way to the Future of...* (Mar. 2023), p. 9794. ISSN: 2416-5999. DOI: 10.46298/jdmdh.9794.
- [50] Roelli, Philipp. “The Corpus Corporum, a new open Latin text repository and tool”. fr. In: *Archivum Latinitatis Medii Aevi* 72, no. 1 (2014), pp. 289–304. ISSN: 0994-8090. DOI: 10.3406/alma.2014.1155.
- [51] Shadrova, Anna. “Topic models do not model topics: epistemological remarks and steps towards best practices”. en. In: *Journal of Data Mining & Digital Humanities 2021* (Oct. 2021), p. 7595. ISSN: 2416-5999. DOI: 10.46298/jdmdh.7595.
- [52] Straka, Milan, Straková, Jana, and Gamba, Federica. “UFAL LatinPipe at EvaLatin 2024: Morphosyntactic Analysis of Latin”. In: (2024). DOI: arXiv:2404.05839.
- [53] Summerlin, Danica. “Using the ‘Old Law’ in Twelfth-Century Decretal Collections”. In: *New Discourses in Medieval Canon Law Research: Challenging the Master Narrative*, ed. by Christof Rolker. Series Title: Medieval Law and Its Practice 28. Leiden/Boston: Brill, 2008, pp. 145–169. ISBN: 978-90-04-24816-8.
- [54] “The patristic legacy to c. 1000”. In: *The New Cambridge History of the Bible*. 1st ed. Cambridge University Press, Apr. 2012, pp. 505–535. ISBN: 978-0-521-86006-2 978-1-139-05055-5. DOI: 10.1017/cho19780521860062.029.
- [55] Tkacz, Catherine Brown. “”Labor Tam Utilis”: The Creation of the Vulgate”. In: *Vigiliae Christianae* 50, no. 1 (1996), p. 42. DOI: 10.2307/1584010. (Visited on 10/28/2025).

- [56] Tolonen, Mikko, Hill, Mark J., Ijaz, Ali Zeeshan, Vaara, Ville, and Lahti, Leo. “Examining the Early Modern Canon: The English Short Title Catalogue and Large-Scale Patterns of Cultural Production”. en. In: *Data Visualization in Enlightenment Literature and Culture*, ed. by Ileana Baird. Cham: Springer International Publishing, 2021, pp. 63–119. ISBN: 978-3-030-54913-8. DOI: 10.1007/978-3-030-54913-8\_3. (Visited on 10/26/2025).
- [57] “Types of Biblical Intertextuality”. In: *Congress Volume Oslo 1998*. Brill, Jan. 2000, pp. 39–44. ISBN: 978-90-04-11598-9. DOI: 10.1163/9789004276055\_005.
- [58] Van Cranenburgh, Andreas and Bod, Rens. “A Data-Oriented Model of Literary Language”. en. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, 2017, pp. 1228–1238. DOI: 10.18653/v1/E17-1115. (Visited on 10/26/2025).
- [59] Wasserschleben, Herman. *Die irische Kanonensammlung*. German. Reprint original Leipzig 1885. Aalen: Scientia Verlag, 1966.
- [60] Werckmeister, Jean. “The Reception of the Church Fathers in Canon Law”. In: *The Reception of the Church Fathers in the West: From the Carolingians to the Maurists*. Vol. 1. Brill: Leiden, 1997, pp. 51–82.
- [61] Wu, Yaru, Bizzoni, Yuri, Moreira, Pascale, and Nielbo, Kristoffer. “Perplexing canon: a study on GPT-based perplexity of canonical and non-canonical literary works”. In: *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*. 2024, pp. 172–184.

## A Corpus Statistics

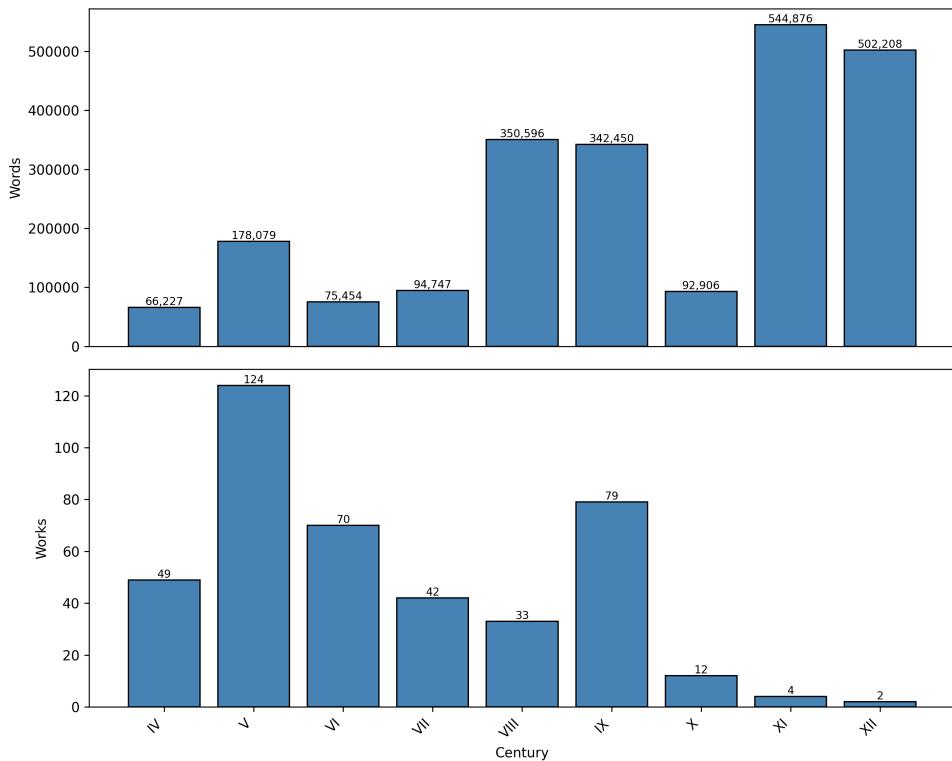
Figure 4 and 5 present chronological distributions of works and word counts per century as well as genres words per genre.

## B Language Modeling

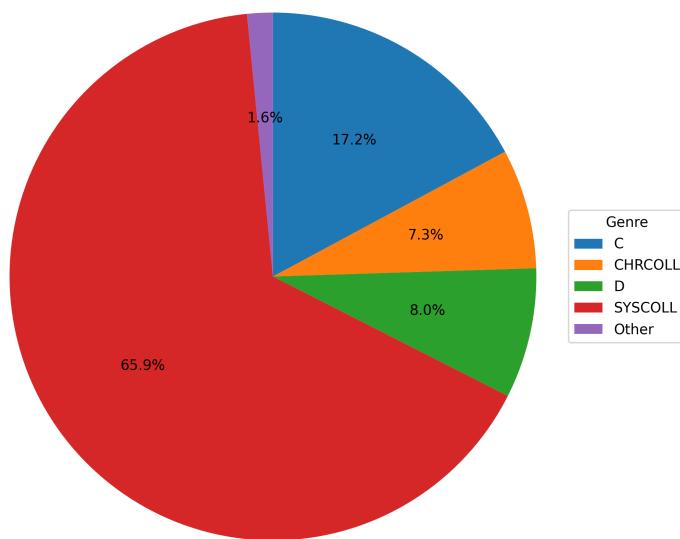
### B.1 GPT-2 Hyperparameters

Parameter	
Layers	6
Attention heads	6
Embedding dimension	384
Dropout	0.2
Vocabulary size	25
Parameters (millions)	10
Learning rate (max)	$1e - 3$
Learning rate (min)	$1e - 4$
Batch	64
Steps	5000
Warmup steps	200

**Table 3:** GPT-2 Hyperparameters and Training



**Figure 4:** Chronological structure of the corpus.



**Figure 5:** Genres represented in the corpus. **SYSCOLL**: systematic collections; **CHRCOLL**: chronological collections; **D**: decretal; **P**: penitentials; **C**: councils.

The model and experimentation framework was implemented in PyTorch Lightning, relying on Andrej Karpathy’s implementation, nanoGPT.<sup>7</sup> All the code is publicly available on GitHub.<sup>8</sup>

## B.2 Preprocessing

For the purposes of character-level language modeling, we created a heavily simplified version of the data by reducing the vocabulary to 25 characters. This was done not only to reduce the model’s size and make it trainable on limited data, but also to mitigate potential discrepancies in typography and orthography arising due to variations in editorial conventions and the inherent instability of medieval spellings. The preprocessing decisions were informed by Ph. Roelli’s recommendations on preparing texts for computer-assisted stemmatological analysis [49]. This simplified version of the data was used exclusively for training language models and computing perplexity. UD-Pipe morpho-syntactic annotation, by contrast, was carried out on significantly less preprocessed versions of the texts.

## C Classification-based Validation

Parameter	
Classifiers	RandomForestClassifier, SVC, KNeighborsClassifier
Cross-validation	StratifiedShuffleSplit
CV-iterations	100
Samples per class per iteration	500
Test size	0.2
Feature space	<i>tf-idf</i> -weighted 2-4 n-grams
Min. document frequency	5
Max. document frequency	0.5
Dimensionality reduction	TruncatedSVD
Components	50

**Table 4:** Classification-based validation parameters.

Table 4 presents parameters of the classification-based validation implemented in `sklearn`. Conceptually similar to the one described [28], our procedure reduces dimensionality and opts for another cross-validation strategy.

## D Standard Score

Model/Corpus	Bible	Canon law	Roman law	Classical Latin	Patristic preaching
<b>Bible</b>	3.59	5.70	6.25	6.36	5.11
<b>Canon law</b>	4.48	3.55	4.49	5.37	4.10
<b>Roman law</b>	5.82	4.49	3.42	5.75	5.33
<b>Classical Latin</b>	4.80	4.28	4.44	4.30	4.32
<b>Patristic preaching</b>	4.00	3.95	4.81	5.15	3.78

**Table 5:** Raw perplexity values for cross-corpus validation (see Table 1).

<sup>7</sup> <https://github.com/karpathy/nanoGPT>.

<sup>8</sup> <https://github.com/Solemne-ERC/canon-law-biblicality>.

Formally,  $z$ -score (or standard score) normalization involves dividing the difference between a raw value and the mean by the standard deviation. In the context of comparative language model evaluation, the rationale for using this normalization is as follows. The raw perplexity values of models trained on small and diverse corpora are influenced not only by the linguistic differences between corpora, which are the primary focus of the evaluation. Other factors beyond our control and interest are the intrinsic complexity of each corpus and the limited capacity of small models to generalize from such scarce data. As a result, raw perplexity values from different models can be thought of as being on different scales, making direct comparison difficult or misleading.

Normalization standardizes each model’s perplexity scores in relation to its performance across all corpora. This allows us to describe the model’s performance on a given corpus in terms of how it deviates from its average performance. Therefore, a low negative  $z$ -score indicates that the model performs better than average (i.e., the corpus is easier for the model to predict). A positive  $z$ -score suggests that the model performs worse than average (i.e., the corpus is more challenging). This approach enables a fairer comparative evaluation and ranks the corpora based on their relative ease or difficulty for any given model.

## E Perplexity-based Measures

### E.1 Perplexity-based Distance

Campos, Gamallo, and Alegria proposed the following perplexity-based distance [9, p. 6]:

Perplexity (PP) is a robust metric to calculate distance between languages. Perplexity measures how well a language model fit the test data. A perplexity-based distance (PLD) between languages or varieties is determined by comparing n-gram based language model (LM) of language L1 and test text (CH) of language L2. The comparison must be made in the two directions.

$$\text{PLD}(L_1, L_2) = \text{PP}(\text{CH}_L_1, \text{LM}_L_2) + \text{PP}(\text{CH}_L_2, \text{LM}_L_1)$$

### E.2 Positive Cross-Score

Konstantinidou, Pavlopoulos, and Barker employed perplexity in what they call a positive cross-score (PCV), defined as “the difference between the PPL for that verse computed with the model trained on the source poem and the equivalent PPL computed with the model trained on the other poem” [28, p. 262].