Umanistica Digitale – ISSN: 2532-8816 – n. 22, 2026: Special Issue
A. Mambelli, L. Bigoni, D. Dainese, F. Tutrone, D. Caffagni, F. Cocchi, M. Zanella, M. Cornia, R. Cucchiara – *The Biblical Heritage in Ancient Latin Christian Literature: Advancing Intertextual Mapping Through Sentence Embeddings*
DOI: http://doi.org/10.60923/issn.2532-8816/22160

The Biblical Heritage in Ancient Latin Christian Literature: Advancing Intertextual Mapping Through Sentence Embeddings

Anna Mambelli

Department of Education and Humanities, University of Modena and Reggio Emilia, Reggio Emilia, Italy – Fondazione per le scienze religiose (FSCIRE), Bologna, Italy

anna.mambelli@unimore.it

Laura Bigoni

Department of History and Cultures, University of Bologna, Bologna, Italy

laura.bigoni4@unibo.it

Davide Dainese

Department of History and Cultures, University of Bologna, Bologna, Italy – Fondazione per le scienze religiose (FSCIRE), Bologna, Italy

davide.dainese@unibo.it

Fabio Tutrone

Department of Cultures and Societies, University of Palermo, Palermo, Italy

fabio.tutrone@unipa.it

Davide Caffagni

Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Modena, Italy

davide.caffagni@unimore.it

Federico Cocchi

Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Modena, Italy – Department of Informatics, University of Pisa, Pisa, Italy

federico.cocchi@unimore.it

Marco Zanella

Department of Mathematics, University of Padua, Padua, Italy

marco.zanella@unipd.it

Marcella Cornia

Department of Education and Humanities, University of Modena and Reggio Emilia, Reggio Emilia, Italy

marcella.cornia@unimore.it

Rita Cucchiara

Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Modena, Italy

rita.cucchiara@unimore.it

Abstract

This study presents an interdisciplinary methodology for detecting biblical references in Latin patristic literature through an innovative combination of rigorous philological approach and Natural Language Processing (NLP) techniques. Focusing on one of the most influential ancient Christian commentaries on the Bible, Augustine of Hippo's *De Genesi ad litteram*, and its relationship with Latin biblical texts (specifically, Jerome's *Vulgate* and pre-*Vulgate* versions), this research introduces a token-based classification system for intertextual references, enriched with semantic annotations and supported by the INCEpTION platform. The first section shows how this numerical classification system accounts for exact matches, lemmatized forms, roots, synonyms, and other forms of semantic parallels (here referred to as "structures"), capturing a wide spectrum of textual similarity. To enhance automatic retrieval of these intertextual connections, we fine-tune BERT-based language models for Latin, incorporating contrastive learning and hard negative mining. In the second section, experimental results show that fine-tuned models significantly outperform baseline models at various levels of textual similarity. This work highlights the utility of computational models in overcoming the traditional dichotomy between explicit quotations and implicit allusions, embracing multiple intermediate nuances of similarity and offering a scalable approach to the study of intertextuality in ancient writings.

Keywords: Latin Bibles, Latin Patristics, Intertextuality, BERT-based Sentence Embeddings, IRCDL2025.

Questo studio presenta una metodologia interdisciplinare per l'individuazione dei riferimenti biblici nella letteratura patristica latina, attraverso un intreccio innovativo di rigore filologico e tecniche di *Natural Language Processing* (NLP). Focalizzandosi su uno dei più significativi commentari cristiani antichi alla Bibbia, il *De Genesi ad litteram* di Agostino d'Ippona, e sul suo rapporto con i testi biblici in latino (in particolare la *Vulgata* di Gerolamo e le versioni precedenti), la ricerca introduce un sistema di classificazione dei riferimenti intertestuali basato su *token*, arricchito da annotazioni semantiche e supportato dalla piattaforma INCEpTION. La prima sezione dell'articolo illustra come questo sistema di classificazione numerica comprenda corrispondenze esatte, forme flesse, radici, sinonimi e altri tipi di parallelismi semantici (qui definiti "strutture"), catturando un ampio spettro di similarità testuale. Per migliorare il recupero automatico di queste connessioni intertestuali, alcuni modelli linguistici per il latino basati su BERT vengono sottoposti a *fine-tuning*, integrando tecniche di *contrastive learning* e *hard negative mining*. Nella seconda sezione, i risultati sperimentali mostrano che i modelli sottoposti a *fine-tuning* ottengono risultati nettamente migliori rispetto ai modelli di base a vari livelli di similarità testuale. Questo lavoro mette in evidenza l'utilità dei modelli computazionali nel superare la tradizionale dicotomia tra citazioni esplicite e allusioni implicite, accogliendo molteplici sfumature intermedie di similarità e offrendo un approccio scalabile allo studio dell'intertestualità negli scritti antichi.

Parole chiave: Bibbie latine, patristica latina, intertestualità, sentence embeddings basati su BERT, IRCDL2025

Introduction

This contribution[1] aims to show how, within the *uBIQUity* project, the intertwining of Humanities and Computer Science methodologies has allowed the team to address complex challenges in the specific field of sentence similarity research within ancient texts.

The *uBIQUity* project, which incorporates the "BI" of the Bible(s) and the "QU" of the Qurʾān in its title, is the Work Package 8 of the larger PNRR project "ITSERR – Italian Strengthening of the ESFRI RI RESILIENCE". The goal of *uBIQUity* is to investigate the sacred texts of Christianity and Islam in different environments and historical periods through two huge *corpora*: Greek and Latin Christian commentaries (broadly understood as written forms of exegesis) on the Bible(s) composed from the patristic age until the late Byzantine period, and classical commentaries on the Qurʾān written in Arabic (*tafāsīr*) from the rise of Islam until the 15th century. These works are unique sources for the study of knowledge, readings and hermeneutics of the sacred texts through the centuries. The intertextual references, conscious or unconscious, that the ancient commentaries contain work as invisible "places of memory", thus making sacred texts "ubiquitous" (hence the title of the project). Indeed, these references, once placed in new contexts for new audiences and readers, become something other than themselves while continuing to refer to themselves and their source text, for those who can still grasp it. Since intertextuality involves the (re)living of a sacred text or tradition in a new and different context, this literary-historical phenomenon also has implications from an exegetical-theological perspective.[2] What the *uBIQUity* project is interested in, therefore, is not pure literary and quantitative data. Rather, this research on *lieux de mémoire* focuses on reconstructing the individual and collective "memories" and the traditions of religious communities over time and space, as well as exploring the recurrent exegetical methods used by the cultured members of these communities. The two aspects are closely connected, as authoritative interpretations of sacred texts shape memories, beliefs and practices within faith communities, and their oscillation can lead to shifts in religious identities.

[1] This paper is the result of a collaborative effort by an interdisciplinary team of philologists, biblical scholars, historians, and computer scientists working within or in collaboration with the *uBIQUity* project. Specifically, the section "A Computable Classification System for Intertextual References" was authored by Anna Mambelli; "Beyond Dichotomy, Reading Between Quotations and Allusions" by Davide Dainese; "The Expansion of the Classification System" by Laura Bigoni; "Memory and Exegesis in Ancient Christian Works" by Fabio Tutrone; the sections "Mapping Intertextuality via BERT-based Models for Latin" and "Experimental Results" by Davide Caffagni, Federico Cocchi, Marco Zanella, Marcella Cornia and Rita Cucchiara. The remaining sections were written collaboratively by the entire team.

[2] This was well illustrated by M. Sternberg [1]. Concerning the use of biblical texts by later authors cf., e.g., the various contributions within the book edited by M.A. Daise and D. Hartman [2] on the reception of the Hebrew Bible in ancient Jewish and early Christian works, or the essays in the volume edited by E.F. Lupieri and L. Painchaud [3] and dedicated to the interpretation of the Apocalypse of John in the light of its use of Hebrew Scriptures.

In this article, we focus on biblical-patristic sentence similarity in Latin, and specifically on the references to Jerome's *Vulgate* and/or pre-*Vulgate* Latin translations of the Bible found in Augustine of Hippo's *De Genesi ad litteram*, one of the most influential works in the history of biblical exegesis. The composition of this commentary situates itself at a crucial stage in the process of stabilization of the Latin biblical tradition, insofar as Augustine wrote it at a time in which different Latin versions of the Bible coexisted and circulated in the Roman empire (Jerome's *Vulgate* and a multifarious universe of pre-*Vulgate* translations, commonly known as *Vetus Latina*). First, we present some case studies related to this ancient commentary on the book of Genesis, as an example enabling us to illustrate the classification system we adopt to manually map the intertextual relationships between the known ancient Greek and Latin versions of the Bible and some patristic texts in the same languages. This token-based manual annotation system represents a shared methodological framework that has been developed in close cooperation with the PRIN 2022 *Resilient Septuagint* team.[3] Through our tagging system carried out on the INCEpTION platform,[4] we aim to overcome the standard dichotomy between unintentional reference ("allusion") and intentional reference ("quotation") to biblical texts that is normally proposed by traditional patristic philology,[5] and to measure the distance between textual passages in a way that better reflects the many nuances and conditions in which intertextual phenomena can appear in ancient works.[6] Although this taxonomy is not a full-fledged ontology in the most popular formats (OWL/RDF etc.) yet, it has allowed us to create a training set to enrich, through human input, the model of automatic numerical representation of the language we train with our own similarity classes. Put another way, this tagging system has been essential for developing

[3] Both *uBIQUity* and *Resilient Septuagint* originate from the methodological framework of the HTLS, edited by E. Bons and D. Scialabba, in collaboration with A. Mambelli [4]. The cooperation between *uBIQUity* and *Resilient Septuagint* has included and will continue to include the development of shared methodological paradigms, interoperable datasets for the study of Greek biblical texts and their heritage, and the prototype of a semantic search engine that can identify biblical references in ancient Greek Christian works with a higher degree of accuracy than pre-existing resources.

[4] INCEpTION is a platform [5] developed between 2017 and 2022 by the Technische Universität Darmstadt and set for *uBIQUity* by the Institute of Science and Technologies of Information (ISTI) of the CNR in Pisa. An environment such as INCEpTION integrates semantic annotation tools with machine learning processes (active learning), providing a simple web interface that is suitable for both our textual sources and the *modus operandi* of computer engineers.

[5] The two main collections of critical editions of Christian texts of the first millennium, the *Corpus Christianorum* by Brepols and the *Sources Chrétiennes* by Éditions du Cerf, classify intertextual references according to this distinction between unintentional and intentional reference: the former is defined as "allusion", tagged with the abbreviation of *confer* (usually "cf."), and the latter as "quotation" (without "cf."). This classification was formalized in 1967 by J. Allenbach [6]. An attempt to expand this classification to include four categories is in *Biblia Patristica*, edited by J. Allenbach, A. Benoît, D.A. Bertrand, *et al.* [7]. For an overview of the different approaches of biblical and patristic scholars to the phenomenon of intertextuality, see, for example, S. Emadi [8].

[6] For a more in-depth methodological reflection on biblical-patristic similarity and for an initial presentation of our token-based classification system of intertextual references, cf. D. Dainese and A. Mambelli [9]; A. Mambelli and M. Costa [50].

the semantic metadata schema adopted for benchmarking the BERT-based Models for Latin that were evaluated [10].

Secondly, we propose fine-tuning existing BERT-based embedding models [11][12][13][14] on annotated Latin *corpora* [15][16][17], using self-generated hard negatives to improve performance in detecting biblical references in ancient Christian literature in Latin. We validate our method through a case study on intertextual analysis in Latin patristic works, in a circular workflow that integrates Natural Language Processing (NLP) techniques with humanistic expertise. This article underscores the transformative potential of interdisciplinary approaches, advancing computational tools for studies on biblical and patristic texts and bridging the gap between philological rigor and computational analysis.

Exploring Biblical References in Latin Patristic Texts

A Computable Classification System for Intertextual References

We begin by illustrating the manual classification system adopted by the *uBIQUity* team to provide a computable account of the intertextual relationships connecting biblical texts and patristic literature. This classification has the following features:

1. It is *token-based*, because it is designed to interact with state-of-the-art textual reuse search systems (*lemma-based*, *n-gram-based*, etc.). According to the methodology of this paper, the term “token” refers to the word, understood as an atomic unit endowed with meaning; for this reason, the two terms are here used interchangeably. The tagging process originates and is performed on the target texts (the Christian commentaries), which means that the token measure applies to them. The source texts (the various ancient versions of the Bible) used for confrontation, on the other hand, are atomized by a different measure, namely the biblical verse. This tagging process aims to distinguish the human approach to texts, which is marked by precise procedures, from the machine approach, which is facilitated by short sequences of characters; at the same time, it attempts to consider the needs of the IT team.

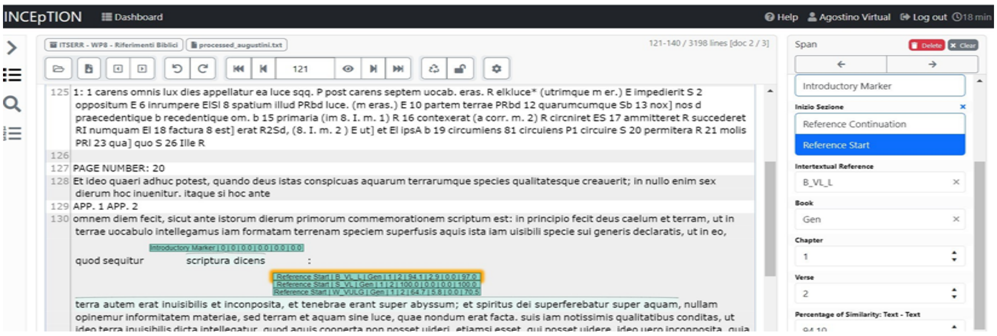
2. Our taxonomy considers, on the one hand, a first type of intertextual reference, in which the words of the ancient Christian commentary coincide totally or partially with some words of the biblical passage identified as source (in one of the biblical versions available in the same language, with the associated apparatus variant readings). These references are assessed for their similarity to the source texts by a numerical proportion based on the number of identical tokens. This type of reference can be handled by classical algorithmics. A second type concerns intertextual references that do not share any identical token with their identified textual sources. These are expected to be assessed by (and to be used as training materials for) Large Language Models, based on analogies of topic and context. These two kinds of intertextual reference can be

considered as the two opposite ends of a wide semantic spectrum that we attempt to explore.

3. In calculating string distance, our method is designed to preserve information concerning variant readings from both direct and indirect textual transmission, which is contained in the apparatus of the editions we use. More precisely, we compare the textual strings of the patristic works in Latin with the three modern reference editions of the ancient Latin Bibles: Jerome's *Vulgate* in the Weber-Gryson edition (W_VULG) [18], and the fragmentary pre-*Vulgate* Latin translations of the Bible in the Sabatier edition (S_VL) [19] and in the Beuron edition (B_VL) [20]. The Weber-Gryson edition presents a reconstructed text of Jerome's *Vulgate* with critical apparatus, showing variant readings from different manuscripts and traditions. The Sabatier edition is the work of the 18th-century Benedictine monk Pierre Sabatier, who gathered a limited number of the *Vetus Latina* manuscripts and combined them with quotations from the Church Fathers to reconstruct the pre-*Vulgate* text of the Latin Bible. Sabatier's edition predates the establishment of the critical philological method in the 19th century and does not include any critical apparatus *sensu proprio*, but as of today, it is the only complete edition of the *Vetus Latina*. The most recent edition of the *Vetus Latina*, produced by the scholarly group of the Vetus Latina-Institut at the Benedictine Archabbey of Beuron, is not yet complete, but for the books already extant, it provides variant readings from many different branches of the tradition. The choice of including variant readings, if available (this is not the case of Sabatier edition), in the calculation of textual similarity reflects an emerging need and philological challenge in the scientific community of Biblical Studies. Indeed, the text-reuse tools such as TRACER, in their current state, generally provide the reconstructed text of only one edition for each ancient work and without a critical apparatus.[7] This flaw of digital resources on the Bible(s) has evident negative consequences from a methodological and qualitative-scientific point of view. Indeed, the biblical tradition has always been intrinsically plural, with different versions of the same texts circulating among different communities over the centuries. The picture of a single text, be it the reproduction of a single manuscript or an eclectic edition, is not suitable for any historical, literary, or exegetical-theological analysis. Therefore, our method of textual annotation illuminates this complexity by embracing multiple ancient biblical versions and placing them in dialogue with the centuries-long tradition of exegetical works.

[7] In addition to TRACER, which was created by Marco Büchler (https://tracer.gitbook.io/manual), see the *Tesserae* project directed by Neil Coffee (https://www.buffalo.edu/digital-scholarship-studio-network/projects/faculty-projects/tesserae.html). However, there are other projects, such as *Musisque Deoque* and *Digital Latin Library*, which experiment with forms of digital visualization of the traditional critical apparatus: cf. S.J. Huskey [21]. In the biblical field, cf. the ongoing ASTAGS project [22].

The functions we envisaged for the annotation work, based on the needs of our specific research domain, were set up in the INCEpTION platform and made visible, with a very intuitive interface.



1: 1 carens omnis lux dies appellatur ea luce sqq. P post carens septem uocab. eras. R elkluce* (utrimque m er.) E impedierit S 2 oppositum E 6 inrumpere ElSl 8 spatium illud PRbd luce. (m eras.) E 10 partem terrae PRbd 12 quarumcumque Sb 13 nox] nos d praecedentique b recedentique om. b 15 primaria (im 8. I. m. 1) R 16 contexerat (a corr. m. 2) R circniret ES 17 ammitteret R succederet RI numquam El 18 factura 8 est] erat R2Sd, (8. I. m. 2) E ut] et El ipsA b 19 circumiens 81 circuiens P1 circuire S 20 permitera R 21 molis PRI 23 qua] quo S 26 Ille R

PAGE NUMBER: 20

Et ideo quaeri adhuc potest, quando deus istas conspicuas aquarum terrarumque species qualitatesque creauerit; in nullo enim sex dierum hoc inuenitur. itaque si hoc ante

APP. 1 APP. 2

omnem diem fecit, sicut ante istorum dierum primorum commemorationem scriptum est: in principio fecit deus caelum et terram, ut in terrae uocabulo intellegamus iam formatam terrenam speciem superfusis aquis ista iam uisibili specie sui generis declaratis, ut in eo,

Introductory Marker | 0 | 0 | 0.0 | 0.0 | 0.0 | 0.0

quod sequitur scriptura dicens :

Reference Start | B_VL_L | Gen | 1 | 2 | 94.1 | 2.9 | 0.0 | 97.0
Reference Start | S_VL | Gen | 1 | 2 | 100.0 | 0.0 | 0.0 | 100.0
Reference Start | W_VULG | Gen | 1 | 2 | 64.7 | 5.8 | 0.0 | 70.5

terra autem erat inuisibilis et inconposita, et tenebrae erant super abyssum; et spiritus dei superferebatur super aquam, nullam opinemur informitatem materiae, sed terram et aquam sine luce, quae nondum erat facta. suis iam notissimis qualitatibus conditas, ut ideo terra inuisibilis dicta intellegatur, quod aquis cooperta non posset uideri, etiamsi esset, qui posset uidere, ideo uero inconposita, quia

Introductory Marker

Inizio Sezione: Reference Continuation / Reference Start

Intertextual Reference: B_VL_L

Book: Gen

Chapter: 1

Verse: 2

Percentage of Similarity: Text - Text: 94.10

Figure 1. Example from the INCEpTION dashboard.

For each biblical reference known in literature or identified by the *uBIQUity* team in Augustine's *De Genesi ad litteram*, we highlight the *Introductory Marker*, if present, namely the formula that sometimes introduces the scriptural reference, making it explicit to readers: for example, *sicut legimus*, *scriptum est*, *dictum est*, *dicente Deo*, *scriptura dicens*. Then, the beginning and end of the biblical reference are highlighted using the functions *Reference Start* and *Reference Continuation*. For each edition of each biblical version placed in *Intertextual Reference* (B_VL, S_VL, W_VULG), we indicate the *Book*, *Chapter* and *Verse* of the biblical passage referred to in the Augustine's commentary. Once all the information about the biblical reference has been entered, the *Percentage of Similarity* between the annotated commentary string and the biblical passage is calculated, by comparing *Text-Text*, *Text-Apparatus*, and *Apparatus-Apparatus*. Specifically, we assign a full score to exact matches between identical tokens in the comparison between the edited texts of the commentary and the biblical version under consideration (*Text-Text*).

Example 1:

Aug., *Gen. ad litt.*, I, 1 (edited by Zycha [23])[8]	S_VL, Gen 1:1
In principio fecit deus caelum et terram. "In the beginning, God made the sky and the earth."	*In principio fecit Deus caelum et terram.* "In the beginning, God made the sky and the earth."

[8] The digital text of this edition was downloaded from the *Corpus Corporum* database, created by the University of Zurich (https://mlat.uzh.ch/), and revised by the *uBIQUity* team members.

In this case, 7 out of 7 words are identical, so the *Percentage of Similarity* in the *Text-Text* field is filled with 100%. Where 100% is reached with the *Text-Text* comparison, the critical apparatus need not be analyzed, and the same percentage should also be repeated in the *Percentage of Similarity: Final Run* field.

Example 2:

Aug., *Gen. ad litt.*, I, 13	W_VULG, Gen 1:2
Terra autem erat inuisibilis *et* inconposita, *et tenebrae* erant *super* abyssum; *et spiritus dei* superferebatur *super* aquam. “And the earth was *invisible* and *formless*, and darkness *was* over the *abyss*; and God’s spirit *was hovering* over the *water*.”	*Terra autem erat* inanis *et* vacua *et tenebrae super* faciem abyssi *et spiritus Dei* ferebatur *super* aquas. “And the earth was *inane* and *void*, and darkness (*was*) over *the surface of the abyss*; and God’s spirit *was lingering* over the *waters*.”

Here, 11 out of 17 words are identical: 64.7% should therefore be entered in the *Percentage of Similarity: Text-Text* field. Since 100% is not achieved in the *Text-Text* intersection, we proceed by examining the critical apparatus of Zycha and Weber-Gryson’s *Vulgate*.

Exact matches between tokens from the edited text and those in the critical apparatus (either of the commentary or the Scripture: *Text-Apparatus*) receive half the score (0.5), since we rely on the work of the editors in establishing the text. For this reason, we give the variant readings contained in the apparatus of biblical editions a lesser weight, while acknowledging the possibility that those variants may have been available to ancient readers and commentators. We also consider the variant readings of Augustine’s text, since they are attested in the tradition and may reflect a different state of the text’s ancient circulation. We take up the second example used above:

Aug., *Gen. ad litt.*, I, 13	W_VULG, Gen 1:2
Terra autem erat inuisibilis *et* inconposita, *et tenebrae* erant *super* abyssum; *et spiritus dei* superferebatur *super* aquam. “And the earth was *invisible* and *formless*, and darkness *was* over the *abyss*; and God’s spirit *was hovering* over the *water*.” In Zycha’s apparatus we find *aquas* which in W_VULG is in the main text.	*Terra autem erat* inanis *et* vacua *et tenebrae super* faciem abyssi *et spiritus Dei* ferebatur *super* aquas. “And the earth was *inane* and *void*, and darkness (*was*) over *the surface of the abyss*; and God’s spirit *was lingering* over the *waters*.” In the apparatus of Weber-Gryson we find + *erant* which in Zycha is printed in the text.

Therefore, in this case *aquas* and *erant* are worth 2.9% (0.5:17 words = X:100), the sum of which is 5.8%. We enter this information in the field *Percentage of Similarity: Text-Apparatus*.

Where the match occurs exclusively between tokens found in the critical apparatus of the commentary and the Scripture, the score is further halved (0.25). Continuing with example 2, let us imagine (although it is not the case) that in Zycha's apparatus there was *informis* in reference to the earth, and that this same variant was also found in the apparatus of Weber-Gryson. This would show that, in a different branch of the tradition, the two texts might have been closer than they appear from the form given to them by the editors. This variant reading would be worth 0.25, so in our case 1.5% (0.25:17 words = X:100) should be entered in the field *Percentage of Similarity: Apparatus-Apparatus*.

Finally, the variant values entered in the *Text-Apparatus* and *Apparatus-Apparatus* fields are summed with the score of the *Text-Text* field, to account for the possibility that the intertextual reference in the commentary traces the biblical passage in a (more) precise and literal way, but on the basis of a different branch of tradition than that printed in the text of the editions. Resuming example 2 one last time, we will add up 64.7% of the *Text-Text* similarity and 5.8% of the *Text-Apparatus* concordance, resulting in a total similarity of 70.5% between Aug., *Gen. ad litt.*, I, 13 (ed. Zycha), and W_VULG, Gen 1:2. If the *informis* variant reading had actually been present in the apparatus, the total percentage would have risen to 72%.

The final scores thus include and highlight the "granularity", that is, the textual complexity and exegetical stratification, of all our *corpora*. At the same time, annotation at the level of exact token matching has its limitations and is not sufficient on its own to reconstruct the various ways in which ancient commentators reused biblical texts. Therefore, we have expanded this classification system, as will be shown in the following paragraphs.

Beyond Dichotomy, Reading Between Quotations and Allusions

In our tagging system, the resulting numerical criterion, expressed as a percentage (*Percentage of Similarity*: *Final Run*), offers a quantitative solution to the conceptual limitations of the two main paradigms in computational linguistics: the classificatory/ontological approach and that of text reuse detection mentioned above.[9] As a concrete example of our classification system, we now present a challenging case along with the tagging solutions we developed. The adopted metric proves effective when comparing two strings, whether of equal or unequal length, provided they can be traced back to two atomic units (e.g., a biblical verse and a clause within a patristic text). Very often, however, the intertextual reference can connect passages that exceed the scope of a single atomic unit. This is precisely the case encountered in relation to the rather generic allusion that Augustine makes, in his *De Genesi ad litteram*, to the episode of the Tower of Babel, an allusion that cannot be confined to a specific verse. In *Gen. ad litt.*, I, 2, while commenting on the *fiat lux* of Gen 11:3, Augustine asks:

[9] See, recently, J. Horstmann, C. Lück, and I. Normann [24]. Previously, R.H. Trillin and S. Quassdorf [25] had already discarded classifications based on types considered classic but not based on formal logic. In general, this is undoubtedly a fruitful approach in terms of both building digital archives and long-term preservation [26][27].

> *qua lingua sonuit ista vox dicente Deo:* fiat lux*, quia nondum erat linguarum diversitas, quae postea facta est in aedificatione turris post diluvium?*
>
> "In what language did this voice resound when God said: *Let there be light*? For the variety of languages was not yet there, which arose later, when the tower was built, after the flood."

The reference to Gen 11:1–10 is evident; however, there is no single verse that allows us to anchor Augustine's question to a uniquely corresponding atomic unit at the semantic level. It is clear that Augustine has in mind both the construction of the Tower of Babel, mentioned explicitly in Gen 11:4, and the confusion of languages described in Gen 11:7. This implies that, semantically, the verses most closely aligned with Augustine's pericope are in fact two.[10] None of these verses, however, contains a token that also appears in Augustine's text, which does not allow for a token-based alignment.

For this initial benchmarking, no linguistic metadata (such as morphological features, shared roots, synonymy and antonymy, etc.) were taken into account; only the recurrence of identical tokens across two different strings was considered. Otherwise, had morphological variation and synonymy been considered, we could have linked *faciamus* and *turrem* from Gen 11:4 to *aedificatio* and *turris* in Augustine's text, respectively. The same would have been possible with *linguarum diversitas* in Augustine and *confundamus linguam* of Gen 11:7. This would have avoided the risk of overlooking a clear allusion to the biblical text. There is however the expression *post diluvium,* which appears both in the *De Genesi ad litteram* and in Genesis (11:10).

Aug., *Gen. ad litt.*, I, 2	W_VULG, Gen 11:10
Linguarum diversitas, quae postea facta est in aedificatione turris post diluvium. "The variety of languages was not yet there, which arose later, when the tower was built, after the flood."	*Hae generationes Sem Sem centum erat annorum quando genuit Arfaxad biennio post diluvium.* "These are the genealogies of Shem: Shem was one hundred years old when he generated Arphaxad, two years after the flood."

Here, another issue seems to arise regarding the determination of *post diluvium*, which recurs at several points in the preceding chapter (Gen 10) as a temporal marker for certain events.[11] The impasse, in this case, is only apparent, since the *post diluvium* in Augustine's text clearly refers to

[10] Gen 11:4: *Et dixerunt venite faciamus* [*aedificemus* in B_VL] *nobis civitatem et turrem cuius culmen pertingat* [*caput erit usque* in B_VL] *ad caelum et celebremus nomen nostrum* [*faciamus nobis nomen* in B_VL] *antequam dividamur in universas terras* [*dispergamur in faciem omnis terrae* in B_VL]; Gen 11:7: *Venite igitur* [*venite* in B_VL] *descendamus et* [*venite et descendentes* in S_VL] *confundamus ibi* [*illic* in B_VL] *linguam* [*linguas* in B_VL] *eorum ut non audiat* [*audient* in S_VL] *unusquisque vocem proximi sui*.

[11] Gen 10:1 in W_VULG: *Hae generationes filiorum Noe Sem Ham Iafeth natique sunt eis filii post diluvium*; Gen 10:32 in W_VULG: *Hae familiae Noe iuxta populos et nationes suas ab his divisae sunt gentes in terra post diluvium*.

that of Gen 11:10, which concludes the account of the descendants of Shem, during whose time the story of Babel unfolds. The presence of the expression *post diluvium* made it possible to annotate Augustine's passage by splitting the tagging into two parts:

1. *linguarum diversitas, quae postea facta est in aedificatione turris*, for which we opted to establish the link with Gen 11:4 (rather than Gen 11:7), as it seemed to us that the idea of the construction of the Tower and the city of Babel, summarized *in nuce* by Gen 11:4 and already anticipated in 11:3, was more prominent throughout the entire passage of Gen 11:1–10;
2. *post diluvium*, referring to Gen 11:10, the final verse that pertains to this episode.

The phrase *linguarum diversitas, quae postea facta est in aedificatione turris* yields a percentage score of 0 across all available versions/editions, since neither the main texts nor the apparatus of W_VULG, B_VL, nor S_VL (whose apparatus is not considered in this case, as it is a reconstructed text based on Augustine's commentaries) contain any of the tokens found in Augustine's text. The usefulness of intertextual references with zero token-level similarity lies in keeping track of less literal references, in order to test the retrieval system's performance. As for *post diluvium*, its similarity score with the text of Gen 11:10 is 100%, both in W_VULG, as previously noted, and in B_VL[12] and S_VL.[13]

It is precisely the cumbersome nature of such solutions, applied to cases that are, after all, relatively common, that prompted the expansion of the tagset. The goal is to move beyond token-level annotation alone, enabling the tagging of references evoked by other linguistic elements or by deeper conceptual affinities.

The Expansion of the Classification System

To address this textual complexity, as far as the annotations are concerned, the team developed an expansion of the classification, which is aimed at addressing textual similarities beyond the presence of identical tokens in the source and target texts. The methodological framework of the expanded semantic annotation rests on a concentric circles approach that progressively enlarges the numerical classification of intertextual references, going from the exact correspondence to a broader semantic frame. The goal is for the annotation to be able to include and classify other degrees of similarity, which allows the team to analyze and annotate more examples of textual reuse, which in turn may lead to include more textual material in the recall of the search engine.

The concentric circles that capture the semantic expansion of the classification are visible in the image (Figure 2), which gives an overview of the added fields.

[12] *Et hae generationes Sem Sem filius centum annorum cum genuit Arfaxat secundo anno post diluvium.*

[13] *Et hae generationes Sem: Sem filius centum annorum cum genuit Arphaxat, secundo anno post diluvium.*

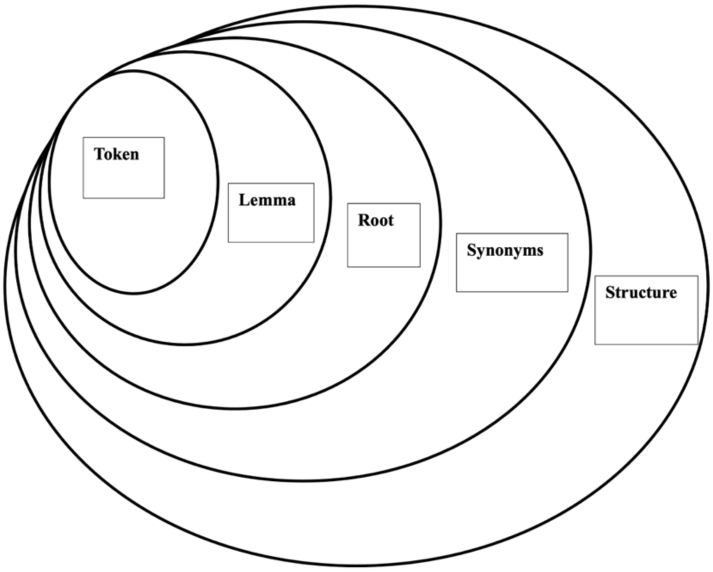


Figure 2. Visualization of the expanded classification.

In order to expand on the representation of the distance between the two texts, we consider those other degrees of similarity, for each of which the calculation is repeated according to the same principle shown above. The tagging is performed at all levels on each string identified as an intertextual reference.

A case of relatively low percentage of similarity obtained by only pairing identical tokens is found at the beginning of Augustine's commentary on Genesis (*Gen. ad litt.*, I, 1). The intertextual reference to Scripture is made clear by the author through the words *secundum id quod Dominus significat, dicens*, "according to what the Lord intends, by saying" that precede it and that are assigned the value of *Introductory Marker*, that highlights recurring expressions of this kind, especially with *verba dicendi* having God (or even Scripture itself) as the subject, as part of the training of the algorithm. In the *incipit* of his work, Augustine introduces a reference to the Gospel of Matthew (13:52) by citing the simile of the scribe who is trained in the kingdom of heaven and is paralleled to the master of a household (*paterfamilias*), who treasures both what is old and what is new; to Augustine, this image refers to the Christian Bible, made of both the First and Second Testaments.

Augustine's passage reads as follows:

> *scribam eruditum in regno Dei similem esse patrifamilias proferenti de thesauro suo nova et vetera.*
>
> "An educated scribe in the kingdom of God is similar to a head of a household who brings out of his treasure what is new and what is old."

Matt 13:52 in turn reads:

> *ait illis ideo omnis scriba doctus in regno caelorum similis est homini patri familias qui profert de thesauro suo nova et vetera.*

> "He said to them: therefore, every scribe who has been trained in the kingdom of the heavens is similar to a head of the household who brings out of his treasure what is new and what is old."

With the first level of classification (Token), 62.5% is calculated by comparing the text of Augustine with both W_VULG and S_VL. Only 9 words out of 15 are identical, hence the abovementioned percentage. The apparatus does not attest to any variant reading and does not contribute to the percentage, which remains low. This could mean that the reference is left undetected by the search engine, or that it may land in a low position in the ordered results. The low percentage in this case does not depend on a concrete shift from the Gospel's text made by Augustine, but rather on the different structure of his phrase as compared to the Gospel. As seen above, the quotation is introduced by a *verbum dicendi*, which transforms the grammatical structure of the quotation into an infinitive clause.

By adding the layer of the lemma, all the tokens that were not found identical in the first step (in this case the remaining 6) are considered. Each token that can be referred to the same lemma as one that is present in the edited text of each biblical verse is given the full value. The proportion is calculated according to the same principles used with the tokens. With this calculation, it is possible to detect 4 more words, that appear in Augustine in a different inflection: *scriba-scribam*, *similis-similem* (nominative-accusative), *est-esse* (indicative-infinitive), *profert-proferenti* (indicative in a relative clause-participle). With this second layer the percentage would be increased to 81.25%.

A further increase is made possible by looking at the fourth level, that of synonyms: all the tokens that have not been classified so far are considered for the Synonym selection (in the example, 2 are remaining, *eruditum* and *Dei*). In this case, the team establishes whether they can be seen as synonymous with one (or more) tokens in the identified source text. The assessment is based on our domain knowledge and linguistic expertise. The case of *eruditum* is simple: *doctus*, which appears in the Gospel, can undoubtedly be considered a synonym. The case of *Dei* is more complex, since it is paired with *caelorum* in Matthew; the two words are not linguistic synonyms, yet they correspond directly in the context. This contextual synonymity is also considered by the team in the process of manual tagging of intertextual references.

With the expanded classification, we are able to map this Gospel citation entirely, thus providing both an analytical approach to Augustine's text in relation to his Biblical sources to the humanist and a numerical and machine-readable tagging to the engineers and software developers.

The last level, called Structure, is the furthest possibility of classification of similarity developed by the team and is based on different principles, since it needs to address questions of intertextual dependence that are not necessarily token-based. For all semantic items (be they tokens, syntagms, syntactical dispositions, metaphors, rhetorical figures, similar topics or other) that have not found a place within the classification so far, we discuss the pertinence of still signaling a similarity that goes beyond those identified by the means of textual semantics listed above [51]. The meticulous classification obtained with this metric seems indicative of the composite nature of intertextual relations between the biblical texts and the ancient authors who made use of it. By assigning different percentages to different degrees of similarity, the picture of a text in relation to its sources appears more nuanced and inspires a scholarly debate on the nature, form, and context of the identified nuances. A classification that may appear rigid at a first glance thus becomes a valuable tool to address different phenomena in textual reuse, which allows scholars to deepen their understanding of the writing habits of an ancient author, going beyond the simple indication of a generic parallelism with the biblical text, which is often all that is found on this matter in the available editions.

Memory and Exegesis in Ancient Christian Works

As shown in the previous paragraphs, Augustine's intertextual references to the multi-faceted tradition of the Latin Bible in his *De Genesi ad litteram* illustrate very well the implicitly transformative (and at times elusive) effects of ancient quotation practices – which, as Antoine Compagnon pointed out, typically provide the reader with both a mutilated organ ("un organe mutilé") and a new, self-sufficient body ("un corps propre, vivant et suffisant"), that is, with a set of older and original texts at the same time [28]. Tracing the origins of the pieces of biblical literature transplanted into the textual body of Augustine's commentary on Genesis can often be a challenging task, especially in light of Augustine's long-standing preference for the extensively ramified universe of the *Vetus Latina*, "a wide variety of translations that have come into existence simultaneously" [29]. The difficulties inherent in the task of identifying the exact branch of the Latin biblical tradition re-used by Augustine become even greater when one considers that, as one of the leading scholars of the *Vetus Latina corpus* observed, "the majority of quotations were made from memory, and in many cases the exact form of text was not important: authors may have paraphrased their biblical source or quoted inaccurately" [30].[14]

It is precisely this rich and complex landscape of mutually overlapping networks that makes the computational approach, intertwined with humanistic expertise, the most suitable for detecting Latin biblical intertextuality. By providing users of the *uBIQUity* tool with scores and data based on varying degrees of similarity, we aim to shed light on often overlooked intertextual relationships without pre-determining the scholars' assessment of the cultural and historical phenomena that underly the ancient Christian interpretation of the Latin Bible: canonization, textual resistance, cognitive distortions, and so on. It is up to the scholars and their critical awareness to interpret the (often nuanced) evidence of textual similarity in one direction or another. One example drawn from Augustine's quotation of a biblical book other than Genesis in his *De Genesi ad litteram* may suffice to clarify this issue.

In Book 1, when commenting on the claim that "the Spirit of God was hovering over the face of the waters" (*Spiritus Dei superferebatur super aquam*),[15] Augustine investigates the theological meaning of the verb *superferre* and its connection with the nature of divine love, which, according to our author, is "a power surpassing and transcending all creatures" (*omnia superante ac praecellente potentia*).[16] In order to corroborate his point, Augustine introduces a quotation from Paul's First Letter to the Corinthians, which, as is typical of Augustine and early Christian

[14] H.A.G. Houghton [30] also remarks on the potentially misleading effects of the changes occurred in the manuscript tradition of patristic works: "The text of biblical quotations may have been adjusted subconsciously, by copyists familiar with a different form, or deliberately, by an editor seeking to bring it into conformity with a current version. Exegetical works such as commentaries are particularly vulnerable to this."

[15] Gen 1:2.

[16] Aug., *Gen. ad litt.*, I, 7, 13. Here and in what follows, translations of Latin texts are original.

commentaries, is carefully integrated into the rhetorical, syntactic, and intellectual structure of the main argument:

> *Cuius rei memor Apostolus dicturus de caritate, supereminentem viam demonstraturum se ait.*
>
> "Remembering this, the Apostle, when he is about to speak on charity, says that he will show the most excellent way".

By mentioning "the Apostle" (*Apostolus*) and the theologically outstanding topic of one of the most well-known sections of 1 Corinthians (*de caritate*, "on charity", which almost certainly echoes a heading of the paratextual tradition of Paul's manuscripts), Augustine takes a characteristically allusive posture that introduces the reader to the content of his Second Testament quotation. Augustine makes thus very clear that he is referring to 1 Cor 12:31, a text central to the life and values of early Christian communities. However, a comparison with the *Vulgate* (W_VULG) and the *Vetus Latina* (S_VL)[17] shows that it is extremely hard, if not impossible, to ascertain in detail which biblical version Augustine is following:

> 1 Cor 12:31 (W_VULG): *aemulamini autem charismata maiora et adhuc excellentiorem viam vobis demonstro.*
>
> "But be eager for the greater graces, and yet I point out to you a still more excellent way."
>
> 1 Cor 12:31 (S_VL): *aemulamini autem dona maiora. Adhuc maiorem viam vobis demonstro.*
>
> "But be eager for the greater gifts. I am still showing you a greater way."

Augustine's definition of charity as "the most excellent way" (*supereminentem viam*) differs crucially from both the *Vulgate* and the *Vetus Latina* traditions in its use of the adjectival participle *supereminentem*. The *Vulgate* reading *excellentiorem* has the advantage of being, like *supereminentem*, a participial form, but remains intrinsically distant (first of all, from a paleographic point of view). Modern readers are likely to be puzzled by the fact that, on the one hand, it has been firmly established that Augustine referred to a *Vetus Latina* version of the Book of Genesis – which John Taylor printed as "the Old Latin Text of Genesis used by Augustine" in an appendix to his English translation of *De Genesis ad litteram* [32] – while, on the other hand, "Augustine came to appreciate the details in Jerome's translation of the New Testament and was happier to use this than the *Vulgate* Old Testament".[18] As happens, one cannot rule out the possibility that Augustine is quoting by memory and is adapting his quotation (consciously or unconsciously) to his exegetical context, for the textual variant *supereminentem*, with the use of the prefix *super*, is significantly close to the verb *superferre*, which, as mentioned earlier, is the

[17] The three volumes of U. Fröhlich [31] do not include a revised Latin text of Paul's epistle, but only a careful study of its transmission.

[18] This citation is taken from P.G. Walsh [33]. Augustine's *Letter* 71, 6 clearly attests that he had a copy of the *Vulgate* Gospels by 403 CE, but this does not mean that Augustine consistently used the *Vulgate* Second Testament after that date.

focus of Augustine’s interpretive efforts in this chapter of *De Genesi ad litteram*. Ultimately, it remains true that learned Church Fathers such as Augustine had their own “mental text of the Bible”.[19]

Therefore, by including in its database several Latin commentaries on the Bible written by different ancient authors, which are already available in open access repositories like *Corpus Corporum*, our digital tool will also help shed light on parallel patristic quotations,[20] contributing to a better understanding of the multiple connections between these different, but clearly interrelated, branches of the ancient Christian heritage.

By offering the scientific community of Religious Studies an easily accessible visualization, in terms of similarity scores, between a target text (Augustine’s commentary) and two possible source texts (the *Vulgate* and the *Vetus Latina*), this search engine will provide a solid basis for this and other similar studies of biblical intertextuality, leaving room for different approaches, aims and conclusions, which may go beyond direct literary dependence and enter the field of memory and exegesis. It is precisely those references that can be identified thanks to deeper conceptual affinities, rather than matching words, that are assigned to the potential of language and sentence embedding models, becoming an integral part of their training material.

Mapping Intertextuality via BERT-based Models for Latin

Building on the detailed annotation framework developed within the *uBIQUity* project, we now turn to the use of Transformer-based language models (specifically, BERT variants) for the automatic identification of intertextual references in Latin patristic texts. The aim is to capture both literal citations and more subtle allusions, especially those that exceed the limits of token-level similarity. This computational approach is designed not as a replacement for the fine-grained manual analysis described above, but rather as a means to scale and generalize the insights gained through it, leveraging Deep Learning-based models to extend the reach of humanistic expertise.

We focus here on Augustine’s *De Genesi ad litteram* as a representative case study, examining its intertextual connections with the Latin Bibles, particularly the *Vulgate* (W_VULG) and the *Vetus Latina corpus* (S_VL). We frame this problem as an information retrieval task: given a query, the objective is to retrieve the most relevant documents from a collection. In our settings, a query q is a passage from Augustine's commentary and documents d are verses from the two considered versions of the Bible. Each query is associated with a positive document d^*, corresponding to an intertextual reference between the commentary and the Bibles. In practice, q may be a literal citation of the biblical verse d^*, or it may just allude to d^*. The former type of relationship is

[19] This definition is from H.A.G. Houghton [34], according to whom “the fact that a phrase is introduced as a quotation is a stronger indication that the preacher is invoking scriptural authority than a direct correspondence with any exemplar.”

[20] On this point, see A. Capone [35], who emphazises the need for “una più ampia analisi dei passi, che tenga presente, oltre ai codici biblici e ai testi di Agostino, anche gli altri scrittori cristiani.”

typically easier to identify by measuring the text overlap between a query and a document. Conversely, allusions to the Bibles are hard to detect, as they require complex semantical analysis, a task that is not trivial even for human experts.

Retrieving Bible Passages from Commentary Sentences

We propose to leverage Transformer-based language models [11], such as BERT [12], to effectively capture the complex intertextual references between patristic commentaries and biblical passages. To this end, let f_θ be a BERT-like pre-trained model. Before processing a query sentence or a document with f_θ, the input is first tokenized. Each token is assigned to a unique integer ID, which acts as an index to select the corresponding embedding in the input embedding matrix of f_θ. This sequence of token embeddings is then passed through a stack of twelve Transformer layers, each comprising two main components: the attention operator, which relates each token to all other tokens in the sequence, and a feed-forward network that processes each token independently. The result is a sequence of output embeddings from the final Transformer layer, one for each token in the input. To obtain a single feature vector (i.e., *embedding*) representing the entire input sequence, we experiment with two aggregation strategies:

- **CLS Token Embedding.** BERT-like models prepend a special classification token (i.e., CLS) to the input sequence. The output embedding corresponding to the CLS is often regarded as a condensed and global representation of the entire input sequence.
- **Token Averaging.** An alternative strategy involves aggregating information from all tokens in the sequence to create a more comprehensive representation. This is achieved by taking the average of the embeddings of all tokens, except the CLS, in the input. Unlike the CLS token, which focuses on providing a global summary, token averaging distributes equal importance to each token, potentially capturing finer-grained information about the input sequence.

These embeddings, representing the query and the document, are mathematically expressed as follows:

$$q = f_\theta(q) \in R^m, \quad d = f_\theta(d) \in R^m.$$

At this point, we measure the relevance of d with respect to q by calculating the cosine similarity between the two vectors:

$$s(q,d) = \frac{q \cdot d}{\|q\|\|d\|}$$

where $\| \cdot \|$ indicates the Euclidean norm. Ideally, the relevance score between a query and its positive document should be maximized. Conversely, the similarity score with respect to any negative document — defined as any document other than the positive one — should be minimized.

Fine-tuning with Self-Hard Negative Mining

While the model f_θ is pre-trained on general language modeling tasks, it has not been specifically trained for the task of text retrieval. To adapt f_θ for this purpose, we fine-tune it using contrastive learning, a method that has proven effective for retrieval and other multimodal tasks [36][37][38]. In detail, given a batch of query-positive document pairs $(q, d^*) \in B$, we embed queries and documents with f_θ, and then we compute the InfoNCE loss function [39]:

$$L = \sum_{(q,d^*)\in B} \frac{exp\left(s(q, d^*)\right)}{exp\left(s(q, d^*)\right) + \sum_{d\in N} exp\left(s(q, d\,)\right)}$$

By minimizing the loss function, we encourage f_θ to map a query and its positive document (q, d^*) to two points on the unit sphere that are close to each other. Conversely, negative documents unrelated to q, that are represented by N in the preceding formula, are pushed away from the embedding representation of q.

Overcoming the Lack of Training Data

A key challenge in training f_θ is the limited availability of commentary queries paired with their corresponding biblical passages. To mitigate this issue, we draw inspiration from self-supervised contrastive learning [40][41] and propose a surrogate task for training. Specifically, we sample a verse from W_VULG as a query q, and pair it with the corresponding verse from S_VL as the positive document d^* (or vice versa). At each training step, we sample N negative documents for each query. In addition, we treat the positive and negative documents from other queries within the same batch as further negatives.

Additional Hard Negative Samples

The previously described procedure, commonly employed in contrastive learning [41][42][43], enhances model sensitivity to the distinctions between related and unrelated documents by exposing it to a larger number of negative samples. The quality of these negatives is crucial: documents that are similar to the query in the embedding space but not semantically related are referred to as *hard negatives*. These hard negatives are known to improve the robustness of models trained with contrastive loss functions [44][45][46][47] like InfoNCE.

In this work, we propose an effective strategy for mining hard negatives during training. First, we generate document embeddings by processing verses from W_VULG with the pre-trained model f_θ. Then, for each positive document d^* associated with a query q, we retrieve the top-k most similar documents and use them as hard negatives for q. Fine-tuning the model using hard negative documents coming from the BERT model itself, as opposed to randomly sampling documents, makes the loss function previously defined more challenging to minimize, ultimately leading to improved performance.

Experimental Results

Experimental Setup

Considered BERT-based Embedding Models

In this study, we model f_θ with three language models sharing the architecture of the BERT model [12], namely Latin RoBERTa [15], Latin BERT [16], and LaBERTa [17]. All considered models have been pre-trained with the masked language modeling objective [12][13]: the model is asked to predict the missing words that are randomly masked in the input sentence. The main difference between the three models is the Latin *corpus* chosen for pre-training. Latin RoBERTa was trained on 390M tokens extracted from the Latin portion of CC-100 [48]. Latin BERT used 642M tokens from a variety of sources spanning the Classical era to the 21st century. Lastly, LaBERTa was trained on *Corpus Corporum*[21] for a total of 167M tokens.

Benchmark Characteristics

Experiments are conducted on 192 annotated references to W_VULG and 170 to S_VL, classified into four similarity categories based on their lexical overlap scores: 0-25%, 25-50%, 50-75%, and 75-100%. As previously described, these similarity ranges capture the spectrum of intertextual relations, from loose thematic connections to *verbatim* citations. Table 1 details the distribution of references across these similarity ranges. Notably, references to W_VULG are distributed relatively evenly, while references to S_VL skew toward high similarity scores, with 83 instances scoring between 75% and 100%. It is also important to note the differing overall sizes of the two biblical *corpora*. W_VULG contains 35,057 passages (each corresponding to a biblical verse), whereas S_VL only comprises 20,791 passages.

		#References				
Corpus	**#Passages**	**0-25%**	**25-50%**	**50-75%**	**75-100%**	**All**
Vulgate	35,057	51	50	46	45	192
Vetus Latina	20,791	44	23	20	83	170

Table 1. Distribution of annotated references across similarity score ranges for the two biblical *corpora*. The total number of biblical passages in each *corpus* is also provided.

Training Details

All models produce embeddings of size m equal to 768. We fine-tune them with the loss function previously detailed, using identical hyperparameters and settings. Specifically, we train with the Adam optimizer [49], a learning rate fixed to 1×10^{-6}, a batch size of 32 queries, and we sample 7 negative documents for each query. Training typically requires 6 hours on a single NVIDIA A40 GPU.

[21] https://mlat.uzh.ch

Impact of Token Aggregation Strategies

Table 2 offers a detailed comparison of the three pre-trained BERT-based models for Latin evaluated in this study, using the W_VULG and S_VL *corpora*. The models are assessed on their ability to retrieve the correct biblical passage for a given query, without any task-specific fine-tuning. Performance is measured using Recall at top-k (R@k) for $k \in \{1, 2, 3, 5, 10\}$. As outlined in the previous sections, the analysis explores two distinct strategies for aggregating token embeddings into fixed-size representations: the CLS token and token averaging.

The results clearly show that token averaging consistently outperforms the CLS token approach, capturing finer-grained information distributed across all tokens in a sequence. This leads to significant performance gains for nearly all models in both *corpora*. Among the three evaluated models, Latin BERT and LaBERTa emerge as the most effective, consistently achieving the highest recall scores and outperforming Latin RoBERTa across most settings. Consequently, the remainder of the paper focuses on Latin BERT and LaBERTa, reporting fine-tuning results using token averaging as the pooling method.

		Corpus: **W_VULG**					***Corpus:*** **S_VL**				
Model	**Pooling**	R@1	R@2	R@3	R@5	R@10	R@1	R@2	R@3	R@5	R@10
Latin RoBERTa	CLS Token	13.0	13.0	13.5	13.5	14.1	4.7	6.5	8.2	8.2	10.0
Latin RoBERTa	Token Avg	**18.2**	**20.3**	**23.4**	**27.6**	**30.2**	**15.9**	**17.6**	**18.2**	**20.6**	**23.5**
Latin BERT	CLS Token	18.2	24.5	26.6	28.1	29.7	18.8	24.1	28.2	32.9	34.1
Latin BERT	Token Avg	**33.3**	**38.0**	**41.7**	**44.8**	**47.9**	**35.3**	**39.4**	**42.9**	**45.3**	**48.8**
LaBERTa	CLS Token	31.3	39.1	**44.3**	47.4	**55.7**	29.4	34.7	**40.0**	**45.3**	**50.6**
LaBERTa	Token Avg	**34.4**	**40.6**	43.8	**47.9**	52.6	**33.5**	**37.6**	**40.0**	43.5	47.6

Table 2. Performance comparison of existing Latin BERT-based models on the W_VULG and S_VL *corpora*, using either the CLS token or the mean of all tokens in the sentence to compute similarities. All results are reported without fine-tuning the embedding model.

Effect of Fine-tuning and Self-Hard Negative Mining

Table 3 compares the performance of Latin BERT and LaBERTa models under various fine-tuning strategies. The results clearly indicate that fine-tuning substantially improves retrieval performance, with the incorporation of hard negatives providing an additional boost, especially in R@1, which is critical for accurate passage retrieval.

Without fine-tuning, both models achieve only moderate performance, with R@1 scores below 35% on both *corpora*. Applying fine-tuning without hard negatives leads to consistent improvements. For example, Latin BERT improves from an R@1 of 33.3% to 38.5% on W_VULG, while LaBERTa increases from 34.4% to 41.1%. Similar improvements are observed on S_VL, along with gains at other recall levels, emphasizing the value of adapting pre-trained models to the specific task of intertextual retrieval.

Introducing hard negatives during fine-tuning further enhances performance across all metrics. Latin BERT shows the largest improvement, with R@1 increasing to 47.4% on W_VULG and 38.8% on S_VL. LaBERTa also achieves significant gains, reaching R@1 scores of 43.2% on W_VULG and 41.8% on S_VL. These results underline the effectiveness of using hard negatives to help models better discriminate between similar and unrelated passages.

		Corpus: **W_VULG**					***Corpus:*** **S_VL**				
Model	**FT**	R@1	R@2	R@3	R@5	R@10	R@1	R@2	R@3	R@5	R@10
Latin BERT	-	33.3	38.0	41.7	44.8	47.9	35.3	39.4	42.9	45.3	48.8
Latin BERT	w/o HN	38.5	46.9	52.1	**55.2**	58.9	35.9	42.3	**47.1**	49.4	54.1
Latin BERT	w/ HN	**47.4**	**51.6**	**54.2**	**55.2**	**59.9**	**38.8**	**42.9**	45.3	**50.0**	**55.3**
LaBERTa	-	34.4	40.6	43.8	47.9	52.6	33.5	37.6	40.0	43.5	47.6
LaBERTa	w/o HN	41.1	50.0	**54.2**	**59.4**	**64.6**	37.1	45.3	**48.8**	**57.6**	**62.4**
LaBERTa	w/ HN	**43.2**	**50.5**	52.1	56.3	63.5	**41.8**	**45.9**	48.2	55.3	61.2

Table 3. Performance comparison of Latin BERT and LaBERTa with different fine-tuning strategies (FT) on the W_VULG and S_VL *corpora*, including results with and without hard negatives (i.e., w/ HN and w/o HN).

Analyzing Performance at Higher Reference Difficulty Levels

Table 4 presents model performance with and without fine-tuning across different difficulty levels, defined by the similarity between a query and its corresponding biblical passage. The lowest similarity range (0-25%) represents the most challenging queries, where there is minimal textual overlap with the target passage. In this range, models without fine-tuning perform poorly, with recall scores nearing zero, highlighting the difficulty of detecting loosely referenced passages. Fine-tuning, however, leads to marked improvements. Notably, LaBERTa achieves a recall of 15.7% on W_VULG and 9.1% on S_VL. As similarity increases (25-50% and 50-75% ranges), performance improves significantly, particularly for fine-tuned Latin BERT and LaBERTa, which attain much higher recall. For example, LaBERTa reaches 69.6% on W_VULG and 40.0% on S_VL. In the highest similarity range (75-100%), all models perform best, with fine-tuned variants of Latin BERT and LaBERTa achieving R@1 scores at or above 70% for both datasets. This analysis indicates that while models are highly effective at identifying close or exact matches, their performance drops significantly with more implicit references, though fine-tuning helps address this limitation.

		0-25%			**25-50%**			**50-75%**			**75-100%**		
Model	**FT**	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Corpus: `W_VULG`													
Latin RoBERTa	✗	0.0	0.0	0.0	8.0	18.0	18.0	34.7	50.0	54.3	33.3	46.7	53.3
Latin BERT	✗	1.9	1.9	5.9	20.0	34.0	38.0	60.8	69.6	73.9	55.5	**80.0**	80.0
LaBERTa	✗	0.0	5.8	9.8	20.0	42.0	48.0	63.0	71.7	78.3	60.0	77.7	80.0
Latin BERT	✓	5.9	15.7	25.5	**40.0**	**52.0**	**52.0**	**73.9**	**78.3**	**82.6**	**75.6**	**80.0**	**84.4**
LaBERTa	✓	**15.7**	**31.4**	**41.2**	26.0	46.0	50.0	69.6	73.9	**82.6**	66.7	77.8	84.4
Corpus: `S_VL`													
Latin RoBERTa	✗	0.0	0.0	0.0	0.0	4.3	4.3	5.0	10.0	15.0	31.3	38.6	42.2
Latin BERT	✗	0.0	2.3	2.3	4.3	8.7	13.0	40.0	45.0	50.0	61.4	78.3	83.1
LaBERTa	✗	0.0	0.0	4.5	8.7	26.1	30.4	**45.0**	50.0	**55.0**	55.4	69.9	73.5

Latin BERT	✓	0.0	6.8	15.9	8.7	17.4	21.7	40.0	45.0	**55.0**	**67.5**	**83.1**	**85.5**
LaBERTa	✓	**9.1**	**27.3**	**40.9**	**13.0**	**34.8**	**34.8**	40.0	**55.0**	**55.0**	**67.5**	75.9	80.7

Table 4. Performance comparison of BERT-based models, with and without fine-tuning (FT), across various subsets of the W_VULG and S_VL *corpora* based on annotated similarity scores.

Conclusion

In this paper, we have shown the efficacy of integrating philological analysis with Transformer-based language models to detect complex intertextual references in Latin patristic literature. The expanded annotation framework developed within the *uBIQUity* project provides a sophisticated classification for intertextual references that moves beyond traditional dichotomy of quotation *versus* allusion. By fine-tuning Latin-specific BERT models with self-hard negative mining, we have improved the retrieval of biblical references across both *verbatim* and semantically distant instances. The results show that semantic enrichment and model adaptation significantly enhance performance, particularly for intertextual references that elude token-based systems. Ultimately, this interdisciplinary approach offers a robust framework for advancing research in textual reuse, biblical studies and exegetical traditions of ancient Christianity.

Acknowledgements

The research presented in this article was mainly supported by the PNRR project "ITSERR – Italian Strengthening of the ESFRI RI RESILIENCE" (Project cod. IR0000014 – CUP B53C22001770006) – National Recovery and Resilience Plan (PNRR) – Mission 4, Component 2, Investment 3.1, funded by the European Union – NextGenerationEU – ref. Notice MUR 3264/2021. The views and opinions expressed, however, are those of the authors and do not necessarily correspond to those of the European Union or the European Commission. Neither the European Union nor the European Commission can be considered responsible for them. In addition to Anna Mambelli, scientific coordinator and product owner of *uBIQUity* (https://www.itserr.it/index.php/wp8-ubiquity/), research team members are Fabrizio D'Avenia (WP8 leader), Sara Abram, Marcello Costa, Cinzia Ferrara, Chiara Palillo and Fabio Tutrone from the University of Palermo. Regarding Computer Science research, for the Latin section, the collaboration with the team from the University of Modena and Reggio Emilia, composed of Davide Caffagni, Federico Cocchi, Marcella Cornia, Rita Cucchiara, and with Cesare Concordia and Carlo Meghini of the National Research Council (CNR-ISTI, Pisa) is fundamental.

The research of Davide Dainese (PI), Laura Bigoni and Marco Zanella is supported by the PRIN (Italian Research Project of National Interest) project "*Resilient Septuagint. An Initial Exploration of the Semantics of Killing and Healing in the Septuagint and its Reception in Patristic and Late*

Antique Sources – 3rd cent. BCE-5th cent. CE” (Prot. 20229E83B3 – CUP J53D23013060001). Other members of the *Resilient Septuagint* team are Luca Arcari (University of Naples Federico II), Laura Carnevale (University of Bari), Isabella Pignocco (University of Catania), Arianna Rotondo (University of Catania), Giorgia Sampò (DDC University of Southern Denmark/University of Bologna) and Gianluca Scatigno (University of Bologna).

References

[1] Sternberg, Meir. 1982. "Proteus in Quotation-Land: Mimesis and the Forms of Reported Discourse." In *Poetics Today* 3 (2): 107–156.

[2] Daise, Michael A., and Dorota Hartman, eds. 2022. *Creative Fidelity, Faithful Creativity: The Reception of Jewish Scripture in Early Judaism and Christianity*. Napoli: UniorPress.

[3] Lupieri, Edmondo F., and Louis Painchaud, eds. 2024. *"Who Is Sitting on Which Beast?" Interpretative Issues in the Book of Revelation*. Turnhout: Brepols.

[4] Bons, Eberhard, and Daniela Scialabba, eds., in collaboration with Anna Mambelli. 2020–. *Historical and Theological Lexicon of the Septuagint* (HTLS). 4 vols. Tübingen: Mohr Siebeck.

[5] Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. "The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation." In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*.

[6] Allenbach, Jean. 1967. *Étapes, moyens et méthode d’analyse pour la constitution du Fichier microphotographique des citations de l’Écriture chez les Pères*. Strasbourg: Université de Strasbourg.

[7] Allenbach, Jean, André Benoît, Daniel A. Bertrand, *et al.*, eds. 1975. *Biblia Patristica: index des citations et allusions bibliques dans la littérature patristique*. 5 vols. Vol. 1, *Des origines à Clément d’Alexandrie et Tertullien*. Paris: CNRS.

[8] Emadi, Samuel. 2015. "Intertextuality in New Testament Scholarship: Significance, Criteria, and the Art of Intertextual Reading." In *Currents in Biblical Research* 14 (1): 8–23.

[9] Dainese, Davide, and Anna Mambelli. 2023–2024. "Intertestualità tra Bibbie e antichi commentari cristiani: l’esempio di *simul* nel *De Genesi ad litteram* di Agostino." In *Lexicon Philosophicum: International Journal for the History of Texts and Ideas* 11: 39–65.

[10] Caffagni, Davide, Federico Cocchi, Anna Mambelli, Fabio Tutrone, Marco Zanella, Marcella Cornia, and Rita Cucchiara. 2025. "Benchmarking BERT-based Models for Latin: A Case Study on Biblical References in Ancient Christian

Literature." In *Proceedings of the 21st Conference on Information and Research Science Connecting to Digital and Library Science*.

[11] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems*.

[12] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

[13] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." *arXiv preprint* arXiv:1907.11692.

[14] Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter." In *Advances in Neural Information Processing Systems*.

[15] Bamman, David, and Patrick J. Burns. 2020. "Latin BERT: A Contextual Language Model for Classical Philology." *arXiv preprint* arXiv:2009.10053.

[16] Ströbel, Patrick B. 2022. "RoBERTa Base Latin Cased v1." https://huggingface.co/pstroe/roberta-base-latin-cased.

[17] Riemenschneider, Frederick, and Anette Frank. 2023. "Exploring Large Language Models for Classical Philology." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

[18] Weber, Robert, and Roger Gryson, eds. ^{5}2007. *Biblia Sacra iuxta Vulgatam Versionem*, Stuttgart: Deutsche Bibelgesellschaft (R. Weber, ^{1}1969).

[19] Sabatier, Pierre, ed. 1743–1751. *Bibliorum Sacrorum latinae versiones antiquae seu Vetus Italica*. 3 vols. Reims: Reginaldus Florentain.

[20] Fischer, Bonifatius, Roger Gryson, Walter Thiele, *et al.*, eds. 1949–. *Vetus Latina: Die Reste der altlateinischen Bibel nach Petrus Sabatier neu gesammelt und herausgegeben von der Erzabtei Beuron*. Freiburg i.B.: Herder.

[21] Huskey, Samuel J. 2019. "The Digital Latin Library: Cataloging and Publishing Critical Editions of Latin Texts." In Monica Berti, ed., *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, 19–34. Berlin-Boston: De Gruyter.

[22] Kauhanen, Tuukka, and Hannu Kalavainen. 2020. "Automated Semantic Tagging of the Göttingen Septuagint Apparatus." In *A Journal of Biblical Textual Criticism* 25: 145–147.

[23] Zycha, Joseph. 1894. *Sancti Aureli Augustini De Genesi ad litteram libri duodecim: eiusdem libri capitula. De Genesi ad litteram imperfectus liber. Locutionum in Heptateuchum libri septem*. Pragae & Vindobonae & Lipsiae: Tempsky & Freyta.

[24] Horstmann, Jan, Christian Lück, and Immanuel Normann. 2023. "Systems of Intertextuality: Towards a Formalization of Text Relations for Manual Annotation and Automated Reasoning." In *Digital Humanities Quarterly* 17 (3): 1–74.

[25] Trillini, Regula Hohl, and Sixta Quassdorf. 2010. "A 'Key to All Quotations'? A Corpus-Based Parameter Model of Intertextuality." In *Literary and Linguistic Computing* 25 (3): 269–286.

[26] Andrews, Tara L., and Caroline Macé, eds. 2014. *Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches*. Turnhout: Brepols.

[27] Tomazzoli, Gaia. 2022. "Intertextuality in Dante's 'Commedia': Hypermedia Dante Network." In *Bibliotheca Dantesca* 5: 308–311.

[28] Compagnon, Antoine. 1979. *La Seconde Main, ou le travail de la citation*. Paris: Éditions du Seuil.

[29] Rose, Paula J. 2013. *A Commentary on Augustine's* De cura pro mortuis gerenda*: Rhetoric in Practice*. Leiden-Boston: Brill.

[30] Houghton, Hugh A.G. 2023. "The Earliest Latin Translations of the Bible." In Hugh A.G. Houghton, ed., *The Oxford Handbook of the Latin Bible*, 6–7. Oxford and New York: Oxford University Press.

[31] Fröhlich, Uwe, ed. 1995–1998. *Epistula ad Corinthios I*. Fasc. 1–3 [*Vetus Latina: Die Reste der altlateinischen Bibel nach Petrus Sabatier neu gesammelt und herausgegeben von der Erzabtei Beuron*]. Freiburg i.B.: Herder.

[32] Taylor, John H., ed. 1982. *St. Augustine: The Literal Meaning of Genesis. Vol. 2, Books 7–12*. Mahwah: Paulist Press.

[33] Walsh, Patrick G., ed. 2017. *Augustine:* De Civitate Dei *(The City of God), Books XIII & XIV*. Liverpool: Liverpool University Press.

[34] Houghton, Hugh A.G. 2008. *Augustine's Text of John: Patristic Citations and Latin Gospel Manuscripts*. Oxford and New York: Oxford University Press.

[35] Capone, Alessandro. 2010. "Review of *Augustine's Text of John: Patristic Citations and Latin Gospel Manuscripts*, by Hugh A.G. Houghton." In *Bryn Mawr Classical Review* 2010.04.29.

[36] Cornia, Marcella, Matteo Stefanini, Lorenzo Baraldi, Massimiliano Corsini, and Rita Cucchiara. 2020. "Explaining Digital Humanities by Aligning Images and Textual Descriptions." In *Pattern Recognition Letters* 129: 166–172.

[37] Sarto, Sara, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. "Positive-Augmented Contrastive Learning for Vision-and-Language Evaluation and Training." *arXiv preprint* arXiv:2410.07336.

[38] Caffagni, Davide, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2025. "Recurrence-Enhanced Vision-and-Language Transformers for Robust Multimodal Document Retrieval." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[39] Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. 2018. "Representation Learning with Contrastive Predictive Coding." *arXiv preprint* arXiv:1807.03748.

[40] Izacard, Gautier, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. "Unsupervised Dense Information Retrieval with Contrastive Learning." In *Transactions on Machine Learning Research*.

[41] Neelakantan, Arvind, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. 2022. "Text and Code Embeddings by Contrastive Pre-Training." *arXiv preprint* arXiv:2201.10005.

[42] Chen, Ting, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. "A Simple Framework for Contrastive Learning of Visual Representations." In *Proceedings of the 37th International Conference on Machine Learning*.

[43] Khosla, Prannay, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. "Supervised Contrastive Learning." In *Advances in Neural Information Processing Systems*.

[44] Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. "Learning Transferable Visual Models From Natural Language Supervision." In *Proceedings of the 38th International Conference on Machine Learning*.

[45] Faghri, Fartash, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. "VSE++: Improving Visual-Semantic Embeddings with Hard Negatives." In *Proceedings of the British Machine Vision Conference 2018*.

[46] Kalantidis, Yannis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. 2020. "Hard Negative Mixing for Contrastive Learning." In *Advances in Neural Information Processing Systems*.

[47] Zhan, Jingtao, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. "Optimizing Dense Retrieval Model Training with Hard Negatives." In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[48] Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. "Unsupervised Cross-lingual Representation Learning at Scale." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

[49] Kingma, Diederik P., and Jimmy Ba. 2015. "Adam: A Method for Stochastic Optimization." In *Proceedings of the 3rd International Conference for Learning Representations*.

[50] Mambelli, Anna, and Marcello Costa. 2025. “Exploring the uBIQUity of Biblical Texts: Tradition and Innovation in the Ancient and Digital Worlds.” In *The Digital Turn in Religious Studies. Research, Services, Infrastructures*. Eds. Alberto Melloni and Francesca Cadeddu. Göttingen: Vandenhoeck & Ruprecht, 149-176.

[51] Dainese, Davide, Laura Bigoni, and Marco Zanella. 2025. “Resilient Septuagint Between Borges and Asimov: A State-of-the-Art Case of Ubiquity.” In *The Digital Turn in Religious Studies. Research, Services, Infrastructures*. Eds. Alberto Melloni and Francesca Cadeddu. Göttingen: Vandenhoeck & Ruprecht, 177-206.