Thematic Dominants of German-Language Catholic Hymns: Automatic Analysis using Word Embeddings

Maria Khokhlova
St Petersburg State University
Russia
m.khokhlova@spbu.ru

Mikhail Koryshev
St Petersburg State University
Russia
m.koryshev@spbu.ru

Abstract

The study analyzes the thematic dominants of German Catholic hymns using methods of distributional semantics. The aim is to automatically identify the semantic structure of a corpus of hymns and to determine its key thematic fields. The material consists of an annotated corpus of German hymns compiled by the authors, including liturgical songs from the „Gotteslob“, from which a list of frequent nouns represented across different editions of the hymnal was extracted. To model semantic relationships and group lexical units, a series of procedures was carried out. First, word embeddings were generated using a multilingual transformer-based model. Second, nonlinear dimensionality reduction was performed using the UMAP algorithm with varying parameter settings (number of neighbors, dimensionality of the space, minimum distance between points, and distance metric). Third, clustering was conducted using the HDBSCAN algorithm, which made it possible to obtain sets of clusters of varying sizes for each model configuration. Model evaluation was carried out using three metrics: the Silhouette Score, the Davies–Bouldin Index, and the Calinski–Harabasz Index. The best-performing model configuration identified 17 thematic clusters of varying sizes, which were subsequently subjected to expert analysis. The findings demonstrate that the semantic space of Catholic hymns is well-structured as a combination of clearly defined thematic cores, which can be grouped into broader categories.

Keywords: word embeddings, German Catholic hymnography, distributional semantics, clustering.

1 Introduction

In recent years, neural network technologies have been increasingly employed in addressing a wide range of linguistic tasks, including the identification of semantically similar words, that is, the detection of groups into which lexical units may be clustered. The previously utilized classical distributional models are based on the frequency characteristics of word co-occurrence; but their principal limitation, however, lies in the fact that the vectors thus generated are, firstly, of high dimensionality and, secondly, characterized by sparsity (owing to zero co-occurrence frequencies) (Landauer and Dumais, 1997).

The subsequent generation of models comprises word vector representations, or embeddings, which likewise enable the capture of semantic relations between lexical units by representing the latter through their contexts, yet are characterized by considerably reduced dimensionality. The transformer architecture introduced in (Vaswani et al., 2017) constitutes one of the most prevalent frameworks in contemporary machine learning and is applied to a broad spectrum of tasks. Thus, the potential applications of the BERT model, which is based upon this architecture, are examined in numerous studies (see, for example, (Gardazi et al., 2025)); a number of investigations have focused specifically on the analysis of particular materials (for instance, the semantic analysis of reviews (Karabašević et al., 2026)). At the same time, deep learning algorithms are employed for the analysis and classification of texts of diverse genres.

A number of studies are devoted to the application of neural network models to the analysis of ecclesiastical chant, for instance, recordings of hymns from the Greek Orthodox Church (Iliadis et al., 2023), to the comparative examination of the Bible, the Qur'an, and the Bhagavad Gita utilizing approaches such as topic modeling and sentiment analysis (Mahit Nandan et al., 2025), as well as to the broader discussion

concerning the potential of processing religious texts by means of computational technologies (Hutchinson, 2024). General issues pertaining to the deployment of artificial intelligence systems within theological inquiry are addressed in (Graves, 2026), whereas (Lima et al., 2025) presents a comprehensive overview of the principal tasks in computational linguistics—including machine translation, authorship attribution, and automatic annotation—that have been undertaken using biblical texts as the primary material.

Although recent studies have demonstrated the usefulness of neural language models for analyzing religious texts, most existing research has focused on text classification, sentiment analysis, topic modeling, or the processing of biblical corpora. Comparatively little attention has been paid to the automatic exploration of semantic structures in hymns. In particular, the identification of thematic domains through distributional semantic representations has not yet been systematically investigated for German-language Catholic hymnography.

To address this gap, the present work studies whether transformer-based word embeddings combined with nonlinear dimensionality reduction and clustering can automatically identify semantically coherent thematic domains in a corpus of German Catholic hymns. The contributions of this paper are threefold. First, the study proposes a pipeline for identifying thematic clusters in a corpus of German Catholic hymns using multilingual contextual embeddings. Second, we evaluate 128 combinations of UMAP (McInnes et al., 2018) parameters using three clustering metrics. Third, we provide an expert interpretation of the resulting semantic clusters and discuss their correspondence to the theological structure of the corpus.

2 Materials and Methods

The corpus used for the study consists of the common (Stammteil) sections of the German Catholic hymnals *Gotteslob* (1975) and *Gotteslob* (2013); the process of its compilation was described in our study (Khokhlova and Koryshev 2024). These hymnals represent the principal official collections of German-language Catholic hymns used in practice. The corpus contains over 41,000 words and includes hymns originating from different historical periods, ranging from medieval texts to contemporary works. Although the corpus reflects the historical development of German Catholic hymnography, the texts were analyzed in their modern editorial form as presented in the two editions of *Gotteslob*. For the purposes of the present study, the corpus was divided into three subcorpora: (1) hymns common to both editions, (2) hymns included only in GL 1975, and (3) hymns introduced in GL 2013.

Since nouns represent the principal part of thematic information in hymnographic discourse, we focused on nouns for further analysis and built a list of lexemes occurring across the three aforementioned subcorpora. After removing duplicate lemmas and manually correcting annotation inconsistencies, the final list comprised 291 nouns (for example, *Gott*, *Leben*, *Jesus*, *Maria*, *Nacht*, *Liebe*, *Herz*, *Erde*, *Tag* and *Zeit*).

The identification of clusters united by a common semantic theme was carried out at the subsequent stage by means of several procedures. Firstly, word vector representations were obtained based on the contexts of word usage within the corpus, with L2 normalization applied so as to transform the vectors to unit length while preserving their direction. To accomplish this task, the pre-trained multilingual transformer-based model paraphrase-multilingual-mpnet-base-v2from the Sentence Transformers library for Python was employed (Reimers and Gurevych, 2019). The selected model provides multilingual sentence and word representations with strong semantic generalization across languages and has demonstrated competitive performance on semantic similarity tasks. Since the corpus contains historical and religious vocabulary, the use of multilingual transformer embeddings was expected to provide more robust semantic representations than static word embeddings.

Secondly, nonlinear dimensionality reduction was performed using the UMAP algorithm, taking into account a range of parameters and facilitated the conduction of a series of experiments. Thirdly, the HDBSCAN algorithm (Campello et al., 2013) was applied for the clustering of lexemes. The formal evaluation of the quality of the model outputs was carried out with reference to the silhouette score (Rousseeuw, 1987), the Davies–Bouldin index (Davies and Bouldin, 1979), and the Calinski–Harabasz index (Caliński and Harabasz, 1974). Since no single clustering metric fully captures cluster quality, three measures were

employed. The silhouette score estimates cluster separation, the Davies–Bouldin index measures cluster compactness and overlap, whereas the Calinski–Harabasz index evaluates the ratio between inter-cluster and intra-cluster variance.

In the course of the investigation, models were constructed with varying parameter settings for the algorithms employed. For the UMAP algorithm, which performs nonlinear dimensionality reduction, the following parameters were taken into consideration:

- Number of neighbors (n_neighbors): 5, 10, 15, 20.

Lower values facilitate the identification of more narrowly defined clusters, whereas higher values enable the description of broader thematic categories.

- Dimensionality of the embedding space, or number of components (n_components): 2, 3, 5, 10.
- Minimum distance between points (min_dist): 0, 0.05, 0.1, 0.15.

This parameter governs cluster density, that is, the extent to which points are positioned in close proximity to or at a distance from one another.

- Distance metric (metric): Euclidean, cosine.

To identify the optimal representation of the semantic space, we performed an exhaustive grid search over the UMAP parameter space. Four values of *n_neighbors*, four values of *min_dist*, four values of *n_components*, and two distance metrics were explored, resulting in 128 parameter configurations, each of which was subsequently evaluated according to the three aforementioned clustering measures.

3 Results

The purpose of the parameter search was not to maximize any single evaluation metric but rather to identify a configuration that provides the best balance between cluster compactness, separation, and interpretability. Consequently, the final model was selected based on the combined behaviour of the three clustering measures together with the semantic coherence of the resulting clusters. The models differ not only with respect to the manner in which lexemes are grouped into clusters and the number of clusters obtained, but also in terms of which lexical units, and in what proportion, were not assigned to any group (i.e., were allocated to the noise cluster, designated by the algorithm as -1). The substantial proportion of such lexemes (which, for individual models, ranges from 40 to 47 percent) may be attributed to the limited size of the corpus employed in the present study.

Table 1 presents the data for the top-10 models that achieved the highest scores according to the aforementioned indices and are ranked by the silhouette metric. The results of the clustering evaluation produced by the models may be interpreted as follows. The silhouette coefficient ranges from -1 to 1: negative values may indicate clustering errors, whereas values approaching 1 imply a clear separation of clusters by the model. Higher values of the Calinski–Harabasz index correspond to models that perform clustering more successfully. For the Davies–Bouldin metric, conversely, lower values are indicative of optimally selected algorithm parameters.

n_neighbours	min_dist	n_components	distance	number of clusters	noise	silhouette score	Davies–Bouldin index	Calinski–Harabasz index
5	0.05	2	cos	17	0.24	**0.60**	**0.50**	**527.57**
20	0.00	3	cos	14	0.41	0.55	0.61	311.60
10	0.10	2	cos	17	0.37	0.55	0.54	408.30
15	0.05	2	euc	17	0.38	0.54	0.53	455.46
5	0.00	5	cos	15	0.18	0.54	0.60	237.80
5	0.00	10	cos	18	0.18	0.54	0.58	216.11
5	0.00	2	euc	16	0.21	0.53	**0.50**	475.74
5	0.00	10	euc	18	0.17	0.53	0.62	218.64
10	0.05	2	cos	16	0.30	0.53	0.53	392.03
10	0.00	3	cos	16	0.28	0.52	0.60	265.72

Table 1: Results for top-10 models.

The highest metric scores were obtained in cases where dimensionality was reduced to two- and three-dimensional spaces (with two exceptions), as well as with the minimum distance between points set to lower values. The majority of the top-10 models employ the cosine measure as the distance metric and distinguish between 15 and 18 clusters. At the same time, the substantial proportion of lexemes designated as noise by Models 2, 3 and 4 merits particular attention.

Model 1 (n_neighbors=5, min_dist=0.05, n_components=2, metric=cosine) proved to be the most successful according to all three evaluation

criteria. It combined a comparatively high Silhouette score with one of the lowest Davies–Bouldin indices while maintaining a high Calinski–Harabasz value resulting in the best overall balance between cluster separation, compactness, and interpretability. More importantly, expert evaluation demonstrated that the resulting clusters were semantically coherent and could be consistently interpreted as thematic domains within the corpus. This model identified 17 clusters, of which the noise cluster accounted for approximately 24% of the lexical units. The latter include, for example, *Sünder, Land, Weg, Angesicht, Bund, Hut, Gericht, Hoffnung, Schmerz, Dorn, Speer* and *Stärke*. The remaining clusters are, on the whole, clearly separated from one another within the resulting embedding space, and semantic homogeneity is likewise observed within them. Although some alternative models achieved slightly better values for individual metrics, Model 1 consistently produced semantically coherent clusters that could be reliably interpreted by experts. It was therefore adopted as the primary model for the subsequent qualitative analysis.

The high performance exhibited by Model 1 may be attributed to the following factors. The selection of a low value for the n_neighbors parameter (equals to 5) enables the UMAP algorithm to preserve the fine-grained local structure of the embedding space, thereby facilitating the identification of compact, semantically homogeneous clusters while avoiding the excessive smoothing that would result in the conflation of distinct themes. The reduction of the embedding space to two dimensions (n_components=2), although entailing some loss of global variance, proves sufficient in this instance to capture the dominant axes of semantic variation within the relatively constrained lexical domain of paraliturgical hymnography. The choice of the cosine distance metric (metric=cosine) appears well-founded: in high-dimensional, sparse vector spaces characteristic of word embeddings, cosine similarity demonstrates greater robustness compared to Euclidean distance. The selected parameter configuration ensures that the resulting clusters reflect genuine contextual synonymy and thematic coherence, rather than constituting artifacts of superficial word co-occurrence.

The subsequent paragraphs examine the remaining clusters in greater detail; thematic labels corresponding to their semantic content have been proposed by the authors for each of these groupings.

1) Marian Emblems and Motherhood: *Mutter, Maria, Kind, Kindlein, Jungfrau, Frau*.

The lexemes pertaining to Mariological themes are, on the whole, successfully delineated, and the inclusion of the lexemes *Kind* and *Kindlein* within this cluster appears harmonious. Nevertheless, the cluster lacks lexemes describing the Nativity thematic complex, which is metonymically associated with the Mariological theme (e.g., *Krippe*, *Stall*).

2) Salvation History in its Temporal Dimension: *Ende, Fürst, Stunde, Anfang, End, Verheißung, Grund, Eva*.

The cluster successfully outlines the temporal perspective of the biblical history of salvation. At first glance, one might expect Adam to appear within this sequence (cf. the doctrine of Christ as the New Adam); however, within the corpus of texts, Adam is mentioned only once, in the hymn “O selger Urgrund allen Seins” (GL 2013, No. 359), which is borrowed from the Liturgy of the Hours.

3) Light and Day Metaphorics: *Stern, Schein, Tag, Licht, Nacht, Finsternis, Morgen, Kyrieleison, Gesicht, Morgenstern, Glanz, Tor, Sonne, Lichte, Aufgang, Morgenröte, Nebel, Mond, Dunkel, Schlaf, Pforte*.

The identification of lexemes associated with the theme of light and day, which subsequently function within metaphorical designations (cf., for example, the image of Mary as the Morning Star in the so-called Litany of Loreto), is complemented by the lexemes *Finsternis*, *Schlaf* and *Dunkel*, which are justifiably assigned to a contrasting cluster based on the principle of opposition. However, this same cluster also includes *Gesicht* ('face'), a lexeme whose connection to this thematic field is merely indirect, as well as *Pforte* ('gate'), which might more appropriately have been assigned to the Mariological cluster.

4) The Realm of Creation as Cosmos and Nature: *Kerub, Feld, Krippe, Pfad, Kreis, Abgrund, Meer, Lauf, Berg, Tal, Ferne, Schnee, Schatten, Paradies, Mantel, Strom, Flut, Wüste, Korn, Wind, Quell, Tier*.

The lexemes describing the created world are delineated with considerable success; even *Kreis* ('circle') appears appropriate, insofar as it expresses the notion of a boundary and the finite completeness of the realm of creation, which stands in opposition to the uncreated celestial world. It is precisely for this reason that *Paradies* ('paradise') ought to be excluded from this cluster, since Paradies in the texts under examination denotes not the created Garden of Eden but rather a locus of the contemplation of the glory of the Almighty.

5) Christ and the Church as the Body of Christ: *Jesus, Christus, Tod, Sterben, Christenheit, Kirche, Todesnot, Retter, Tote*.

This cluster raises no particular objections; the sole refinement that may be warranted concerns the lexeme *Retter* ('savior'), which would more appropriately be assigned to the cluster united by the notion of the messianic dignity of Christ.

6) Eucharistic Meal and Sacrifice Metaphorics: *Fleisch, Abendmahl, Blut, Gast, Mahl, Tisch, Speise, Hunger, Osterlamm, Herd, Wein, Brot, Opfer, Trank, Kelch, Bild, Traube*.

On the whole, this cluster is semantically coherent, uniting lexemes that express the notion of Eucharistic sacrifice and meal. The inclusion of the lexeme *Bild* ('picture, image') within this cluster is, however, infelicitous, given that the concept of the figure, or image, continually accompanies ecclesiastical teaching on the matter of the sacrament as the visible image of a supernatural, transcendent content.

7) Passion, Suffering, Distress, Penance and Sin: *Knie, Schoß, Kreuz, Spott, Stall, Last, Joch, Betlehem, Grab, Hölle*.

This cluster successfully unites lexemes that delineate the notion of Christ's death as a sacrifice for the sins of the world (cf. the versicle from the Good Friday liturgy: “Ecce lignum Crucis, in quo salus mundi pependit” — “Behold the wood of the Cross, on which hung the Savior of the world”) and the idea of penitential adoration as the path toward the reception of this salvation (cf. the responsory to the aforementioned versicle: “Venite, adoremus” — "Come, let us adore").

8) The Kingship of Christ and Messianic Titles: *Himmel, Sohn, Himmelreich, Thron, Volk, Sieg, Heiland, König, Hirt, Reich, Name, Lamm, Stamm, David, Held, Tempel, Himmelsbrot, Hirte, Herrscher, Friede, Josef, Ebenbild, Sieger, Jünger*.

The lexemes conveying the notion of Christ as the reigning Messiah, King of the Universe, are correctly identified. The sole observation that may be made concerns the proper name *Josef*, given that the hymns in question refer not to the Old Testament figure of Joseph the All-Comely—the son of Jacob and a prefiguration of Christ—but rather to Joseph, the Betrothed of the Virgin Mary.

9) Praise and Heavenly Joy: *Lob, Preis, Halleluja, Engel, Dank, Herrlichkeit, Wunder, Freude, Freund, Freud, Wonne, Schar, Gut, Seele, Ruhm, Lobgesang, Zebaot, Gesang, Glück, Loblied, Ehre, Lied, Schöpfung, Staunen, Jubel*.

The set of lexemes included within this cluster raises no objections, with the exception of the lexeme *Schöpfung* ('creation'), which ought to be reassigned to the fourth cluster, the latter comprising lexemes that thematize concepts pertaining to the created world. It should be noted, however, that the lexeme *Schöpfung* is actively employed in constituting contexts of doxological praise addressed to God (cf. Ps 150:6 (Einheitsübersetzung): “Alles, was atmet, lobe den Herrn” — "Let everything that breathes praise the Lord").

10) Faith, Truth, and Spiritual Knowledge: *Macht, Wahrheit, Sinn, Flügel, Weise, Schall, Verstand, Glaube, Geheimnis*.

From the standpoint of Christian theology, the cluster that unites faith, rational cognition, and sensory perception is delineated with considerable success (cf. the Augustinian notion, articulated in his Tractates on the Gospel of John, concerning the act of faith as a necessary step toward the act of genuine knowledge).

11) Prayer, Trust and Divine Help: *Güte, Huld, Heil, Zeichen, Hand, Erbarmen, Treue, Güt, Schutz, Geduld, Demut, Erlöser, Erlösung, Maß*.

This cluster may be merged with the preceding one, given that it merely elucidates particular aspects of an individual's “standing” before God while seeking His countenance (see above).

12) Grace, Consolation and Divine Care: *Trost, Gnade, Dienst, Segen*.

This is a small yet notably “pure” cluster, whose constituent lexemes may likewise be assigned to the thematic field uniting words that describe the relationship between God and humankind. This field encompasses both designations of human actions and deliberately cultivated virtues within this relationship (e.g., *Dienst* 'service, ministry') and the fruits thereof

(e.g., *Gnade* 'grace', *Segen* 'blessing', *Trost* 'consolation').

13) Divine Justice and Judgment: *Gerechtigkeit, Recht, Menschenkind, Urteil, Antwort*.

The clustering procedure successfully recovered the notion of divine justice—specifically, the inevitability of human accountability before God. The grouping is semantically coherent and reflects the theological opposition between justice and mercy frequently discussed in Christian doctrine.

14) Threat, Temptation, and Evil: *Feind, Gefahr, List, Gesetz, Streit, Böse, Plage*.

This cluster unites lexemes designating the forces of evil, their attributes, and the perils that confront the individual along the path of life. The noun *Gesetz* ('law') thematizes, within this cluster, its secondary status for the believer in the context of the Apostle Paul's polemic against proponents of a theology of the law, as opposed to the New Testament conception of salvation by grace.

15) Suffering, Distress and Existential Anxiety: *Pein, Not, Verlangen, Elend, Krankheit, Arm, Angst, Furcht*.

This cluster unites lexemes that delineate the notion of existential dread (*Angst*) experienced by the believer in the face of suffering—a believer who is filled with longing (*Verlangen*) in anticipation of the saving action of the Almighty's right hand (Arm*)* and with reverent, filial awe and veneration (*Furcht*) toward Him.

16) Action, Will and Moral Conduct: *Tun, Rat, Werk, Wille, Mut, Eigentum, Arbeit, Buße, Plan*.

The obtained semantic grouping comprises lexemes that designate being as a purposive (*Plan* 'plan, intention') activity (*Handeln* 'action, agency'), encompassing both volitional (*Wille* 'will', *Mut* 'courage, fortitude of spirit') and moral-ethical aspects (*Buße* 'repentance, penance'), the latter conceived as an attribute of the readiness to assume responsibility for errors committed, to rectify them, and to continue along the path of life in search of the paradise lost by our first parents.

17) Anthropology of Perception and Divine Communication: *Herz, Gnad, Wort, Stimme, Ohr, Mund, Ruf, Zeugnis, Seite, Auge, Hauch, Schweigen, Zunge, Frage*.

This field encompasses lexemes that designate the conceptual sphere of New Testament theological anthropology of perception in its communicative dimension. Within this framework, contextual synonyms include *Wort* ('word') and *Zeugnis* ('testimony'), as well as *Herz* ('heart') and *Gnade* ('grace'). The significance of listening is underscored in contradistinction to speaking, as evidenced by the inclusion of such lexemes as *Schweigen* ('silence'), *Ohr* ('ear'), *Frage* ('question') or *Mund* ('mouth').

The resulting clusters exhibit a high degree of correspondence with thematic domains traditionally distinguished by theological scholarship, suggesting that the proposed workflow successfully recovers meaningful semantic structures without predefined thematic labels. On the whole, a clearly defined cluster structure of the thematic fields within the hymns may be observed, a finding corroborated by the analysis of the results obtained from alternative models. The aggregation into clusters is not of a merely formal character; rather, semantic intersections are discernible in the outcomes.

To evaluate the robustness of the obtained thematic structure, we compared the selected model with two alternative high-ranking configurations (Models 7 and 9). Although these models recovered many of the same thematic domains, they produced either less coherent semantic groupings or a higher proportion of isolated lexical items. This observation suggests that the identified thematic structure is largely stable with respect to moderate changes in UMAP parameters. Models 7 and 9 nevertheless identified several meaningful thematic groups, including “Faith and Cognition”, “Christ the King”, and “Sin and Repentance”. However, the overall thematic organization produced by Model 1 proved to be the most consistent and interpretable. Some clusters correspond directly to established theological concepts (e.g., Mariology or Eucharistic theology), whereas others represent broader conceptual associations that emerge only through the interaction of several lexical items. It may be observed that the clusters are heterogeneous with respect to the number of constituent units they contain; nevertheless, they may, in turn, be aggregated into broader thematic fields.

Overall, the experiments demonstrate that transformer-based semantic embeddings combined with nonlinear dimensionality reduction and clustering constitute an effective approach to exploring the thematic organization of historical religious corpora. The automatically identified

clusters largely correspond to expert interpretations while also revealing previously unnoticed semantic associations, thereby illustrating the potential of distributional semantics for further research.

4 Conclusion

The findings obtained have confirmed the feasibility of automatic cluster identification based on vector representations for religious texts; these clusters, however, necessitate further analysis. The conducted study has demonstrated that the semantic space of German-language Catholic hymns is organized as a set of thematically dense clusters which are, on the one hand, surrounded by a substantial periphery of lexical units lacking robust connections, and, on the other, subsumed within larger groupings. At the same time, the difficulty encountered in assigning thematic labels to specific lexeme groups indicates that, firstly, a comparison with results obtainable through alternative models is required, and secondly, the incorporation of a larger corpus of a genre-similar nature is necessary to achieve greater precision. Thus, the applied approach made it possible to identify clusters and outline the range of themes addressed in the hymns.

References

Tadeusz Caliński and Joachim Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1): 1–27.

Ricardo J.G.B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining (PAKDD 2013)*, volume 7819 of *Lecture Notes in Computer Science*, pages 160–172. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14

D. L. Davies and D. W. Bouldin. 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

Nadia Mushtaq Gardazi, Ali Daud, Muhammad Kamran Malik, Amal Bukhari, Tariq Alsahfi, and Bader Alshemaimri. 2025. BERT applications in natural language processing: a review. *Artificial Intelligence Review*, 58: 166. https://doi.org/10.1007/s10462-025-11162-5

Mark Graves. 2026. Generative AI and theology: a three-year retrospective. *Theology and Science*, 24(1): 1–7. https://doi.org/10.1080/14746700.2025.2592326

Ben Hutchinson. 2024. Modeling the sacred: considerations when using religious texts in natural language processing. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1029–1043. Mexico City.

Lazaros Alexios Iliadis, Sotirios P. Sotiroudis, Nikolaos Tsakatanis, Achilles D. Boursianis, Konstantinos-Iraklis D. Kokkinidis, George K. Karagiannidis, and Sotirios K. Goudos. 2023. Low complexity deep learning framework for Greek Orthodox Church hymns classification. *Applied Sciences*, 13(15): 8638. https://doi.org/10.3390/app13158638

Darjan Karabašević, Aleksandra Vujko, Vuk Mirčetić, Gabrijela Popović, and Dragiša Stanujkić. 2026. A transformer-based semantic encoding framework for quantitative analysis of large-scale textual reviews. *Axioms*, 15(3): 175. https://doi.org/10.3390/axioms15030175

T. K. Landauer and S. T. Dumais. 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104: 211–240.

Bruno Cesar Lima, Nizam Omar, Israel Avansi and Leandro Nunes de Castro. 2025. Artificial intelligence applied to the analysis of biblical scriptures: a systematic review. *Analytics*, 4(2): 13. https://doi.org/10.3390/analytics4020013

Leland McInnes, John Healy, and James Melville. 2018. UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*. https://arxiv.org/abs/1802.03426

Gotteslob. 1975. *Katholisches Gebet- und Gesangbuch*. Ausgabe für das Bistum Trier. Trier: Paulinus Verlag. 1054 S.

Gotteslob. 2013. *Katholisches Gebet- und Gesangbuch*. Ausgabe für das Bistum Trier. Trier. 1296 S.

Maria Khokhlova and Mikhail Koryshev. 2024. A corpus of liturgical texts in German: Towards multilevel text annotation. In *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024)*, pages 201–205. Bulgarian Academy of Sciences.

Mahit Nandan A. D., Ishan Godbole, Pranav M Kapparad, and Shrutilipi Bhattacharjee. 2025. Comparative analysis of religious texts: NLP approaches to the Bible, Quran, and Bhagavad Gita. In *Proceedings of the New Horizons in Computational Linguistics for Religious Texts*, pages 1–10. Association for Computational Linguistics, Abu Dhabi.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Hong Kong, China.

Peter J. Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20: 53–65.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS 2017)*, pages 5998–6008.