



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

Maria Khokhlova, Mikhail Koryshev

St Petersburg State University
m.khokhlova@spbu.ru, m.koryshev@spbu.ru

1. Motivation and Research Questions

Catholic hymns are an important part of European cultural and religious heritage. They convey key concepts of Christian faith and reflect the spiritual, cultural, and historical experience of different epochs. Their thematic structure has not yet been studied systematically with computational methods.

RQ1

What are the main thematic domains of German-language Catholic hymns?

RQ2

How can word embeddings and clustering methods be used to automatically identify and interpret these thematic domains?

2. Data

German hymns from the *Gotteslob* (1975, 2013), *Stammteil*

27,000+

tokens

291

frequent nouns
(lemmas)

Metadata (for each hymn):

- period of origin (7 time bins)
- year of the hymn
- accompaniment epoch
- link to ancestor hymn
- stanza information

3. Methodology

1

Word embeddings

Multilingual transformer-based model
(paraphrase-multilingual-mpnet-base-v2)

2

Dimensionality reduction

UMAP

Varying parameters: number of neighbors,
dimensionality, minimum distance,
distance metric

3

Clustering

HDBSCAN

Clusters of varying sizes

4

Model evaluation

Silhouette Score | Davies–Bouldin Index |
Calinski–Harabasz Index

4. Results

4.1 Best model configuration

UMAP (n_neighbors = 5,
min_dist = 0.05, cosine, 2D)
+ HDBSCAN

17

thematic clusters

≈ 24%

noise points

Evaluation scores

(for the selected model):

Silhouette score	0.60
Davies–Bouldin index	0.50
Calinski–Harabasz index	527.57

4.2 Thematic clusters (based on the article)

#	Theme (interpretation)	Example nouns (German)	#	Theme (interpretation)	Example nouns (German)
1	Marian Emblems and Motherhood	Mutter, Maria, Kind, Kindlein, Jungfrau, Frau	10	Faith, Truth, and Spiritual Knowledge	Macht, Wahrheit, Sinn, Flügel, Weise, Schall, ...
2	Salvation History in its Temporal Dimension	Ende, Fürst, Stunde, Anfang, End, Verheißung, Grund, Eva	11	Prayer, Trust and Divine Help	Güte, Huld, Heil, Zeichen, Hand, Erbarmen, ...
3	Light and Day Metaphorics	Stern, Schein, Tag, Licht, Nacht, Finsternis, Morgen, ...	12	Grace, Consolation and Divine Care	Trost, Gnade, Dienst, Segen
4	The Realm of Creation as Cosmos and Nature	Kerub, Feld, Krippe, Pfad, Kreis, Abgrund, Meer, ...	13	Divine Justice and Judgment	Gerechtigkeit, Recht, Menschenkind, Urteil, ...
5	Christ and the Church as the Body of Christ	Jesus, Christus, Tod, Sterben, Christenheit, Kirche, ...	14	Threat, Temptation, and Evil	Feind, Gefahr, List, Gesetz, Streit, Böse, ...
6	Eucharistic Meal and Sacrifice Metaphorics	Fleisch, Abendmahl, Blut, Gast, Mahl, Tisch, Speise, ...	15	Suffering, Distress and Existential Anxiety	Pein, Not, Verlangen, Elend, Krankheit, Arm, ...
7	Passion, Suffering, Distress, Penance and Sin	Knie, Schoß, Kreuz, Spott, Stall, Last, Joch, ...	16	Action, Will and Moral Conduct	Tun, Rat, Werk, Wille, Mut, Eigentum, Arbeit, ...
8	The Kingship of Christ and Messianic Titles	Himmel, Sohn, Himmelreich, Thron, Volk, Sieg, ...	17	Anthropology of Perception and Divine Communication	Herz, Gnad, Wort, Stimme, Ohr, Mund, Ruf, ...
9	Praise and Heavenly Joy	Lob, Preis, Halleluja, Engel, Dank, Herrlichkeit, ...			

Main findings

- The semantic space of German Catholic hymns is well-structured and consists of clearly defined thematic domains.
- The most prominent clusters are related to Gott and Anbetung, Kirche und Gemeinschaft, Mutter und Kind, Schöpfung und Natur, and Leben, Tod und Ewigkeit.
- A substantial number of nouns form a peripheral group (noise), which reflects the semantic diversity of the corpus.
- The obtained clusters are coherent and interpretable, and correspond to the main theological and cultural concepts represented in the hymns.

5. Discussion

- Word embeddings combined with UMAP and HDBSCAN effectively capture semantic relations between nouns and reveal meaningful thematic clusters.
- The identified thematic domains reflect central Christian concepts as well as themes related to human experience and the natural world.
- The approach enables a large-scale, data-driven analysis of hymn vocabulary and can complement traditional, qualitative methods.
- The results are consistent with the historical and cultural background of the German Catholic hymn tradition.

6. Conclusion

1

The study automatically identifies 17 thematic clusters in a corpus of German Catholic hymns using word embeddings and clustering methods.

2

The thematic structure of the hymns includes both core religious themes and broader human and natural domains.

3

The proposed approach can be applied to other religious and historical text corpora and contributes to the digital humanities and computational research on religious texts.