A Corpus of Liturgical Texts in German: Towards Multilevel Text Annotation

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

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

Abstract

The aim of the study is to create a "documented" literary and theological history of German Catholic hymnography. The paper focuses on the creation of a corpus of liturgical texts in German and describes the first stage of annotation dealing with the metatextual markup of Catholic hymns. The authors dwell in detail on the parameters of the multi-level classification of hymn texts they developed, which allows them to differentiate hymns on different grounds. The parameters include not only characteristics that represent hymns (the period and the source of their origin, rubrics, musical accompaniment), but also ones that are inherent for strophes. Based on the created markup, it is possible to trace general trends in texts divided according to certain meta-features. The developed scheme of annotation is given on the example of the hymnbook Gotteslob (1975). The results present statistics on different parameters used for hymn description.

Keywords: corpus of hymns, liturgy, Catholicism, German language.

1 Introduction

It would not be an exaggeration to say that modern works related to the analysis of language and literature deal with corpora (being either corpus-based or corpus-driven), which have become a necessary condition for such studies. However, most corpora, regardless of the language they are aimed at, are focused on modern language (most often literary), since automatic processing of texts from other periods can still be difficult. Additional difficulties in building corpora of historical texts are associated with the fact that there is no access to the texts in electronic form, so the task of searching, scanning and recognizing them still arises.

In the case of church texts, there is a centuries-old tradition of their collection; however, there are no traditions of their systematization and presentation for research purposes. There are various indexes and reference books, archival materials, but they are fragmented and difficult to access. For example, archival materials are available in the archive of the German Liturgical Institute (Deutsches Liturgisches Institut) in Germany. Therefore, our study makes the first attempt to create a corpus of texts that would include such material.

In the article, we will focus on issues related to the creation of a corpus of liturgical texts using texts from Gotteslob (1975), namely, their metatextual markup. In our case, the created corpus can be classified as both historical and poetic one; therefore metatextual parameters combine the features of these two types. We rely on the principles that were elaborated for the development of similar corpora and TEI project and introduce new ones that take into account the features of liturgical texts and have not previously been presented in online systems. In our work, we limit ourselves to the material from Gotteslob (1975), but we also consistently take into account Gotteslob (2013) and the sources of hymn material used by the editorial boards of the respective editions. In further research, we would also like to consider the regional parts of these editions.

The users of the corpus will be historians and theorists of literature, verse scholars, liturgists, ethnographers, musicologists, and cultural researchers. It is this wide range of different specialists and their scientific interests that influenced the selection of features for meta-annotation.

2 Related work

The scheme of annotation for corpus data varies depending on the types of texts themselves, as well as on the goals that researchers set for themselves. To the best of our knowledge, there are not many projects on presenting poetic texts in electronic form.

The "Deutsch Diachron Digital" project brought together researchers from 15 universities and research institutes in Germany to create a number of diachronic corpora of the German language that cover different periods of its development: from Old German (Altdeutsch) to Early Modern High German (Frühneuhochdeutsch) (Deutsch Diachron Digital).

The Deutsches Textarchiv (DTA) consists of two parts: the main corpus called DTA-Kernkorpus, which includes texts from 1598 to 1913, and the expanded corpus called DTA-Erweiterungen, which covers the period from 1465 to 1969. The first one is balanced and can serve as a reference corpus for studying New High German. The total volume is about 400 million tokens.

The PO-EMO corpus (Haider et al., 2020) contains German and English poetic texts with annotation for esthetic emotions. The mentioned interdisciplinary project is carried out in line with computational poetry research using methods that are implemented in sentiment analysis. The XML markup was done line by line and includes the following parameters: meter, caesura_rhythm, main_accents, caesuras. The corpus of German poetry in New High German (Deutsches Lyrik Korpus, DLK) is described in (Haider and Eger, 2019; Haider, 2021). It includes poetic texts from publicly available German language corpora: the German Text Archive (DTA) and the Digital Library of Textgrid.

The Berliner Repertorium is an interdisciplinary database containing medieval German translations of Latin liturgical songs from the 9th to early 16th centuries (Berliner Repertorium). It comprises 471 Latin hymns, sequences and antiphons with their 3066 Middle High and Low German prose and poetic translations, paraphrases and glosses. Descriptions of the texts are supplemented by digitized copies, which can be useful for further research.

There are no specialized corpora of German-language liturgical poetry, which can be partly explained by the relatively satisfactory documentation of these texts for literary purposes in the early stages of the formation of the genre thanks to the efforts of 19th-century scholars. At the same time, the collection of the Mainz hymnological archive "Mainzer Gesangbucharchiv", counting more than 8,000 units, is at the initial stage of digitization (less than a hundred collections of chants, mainly from the 18th century, have been digitized); the creation of a corpus is not currently planned.

Among the poetic corpora for other languages, one can name the poetic corpus of the Russian National Corpus. It is one of the few that contains poetic texts with a total volume of more than 13 million words. The project "Music and Language in Danish Reformation Hymns" (Svendsen, Sørensen, Troelsgård, 2020) focuses on presenting the Reformation hymns in Danish and developing dictionaries of the corresponding period, which present vocabulary from the hymnbooks.

3 Corpus of liturgical texts in German

The hymns from Gotteslob (1975) cover the period from the Middle Ages to 1973; in the future, the corpus will be supplemented with the results of the analysis of Gotteslob (2013). Text processing involves five stages: preparation, meta-marking, tokenization, morphological and syntactic analysis. The preparatory stage involved determining the list of texts that should be included in the corpus. We have selected hymns that belong to the genre of Kirchenlied (church chants, which are the core part of the hymns). Outside the scope of our attention were the official translations of original Latin prayers, psalms, as well as litanies, texts of formulas, sacramentals and sacraments set to music.

Text mark-up was performed manually using the close reading method, since not all the information about the hymns that should have been presented in the corpus was indicated explicitly in the texts themselves, so they needed to be additionally read by an expert.

4 Text annotation

Text annotation was described in detail in (Sinclair, Ball, 1996), in which the authors identified internal and external markup parameters. The paper by Haider et al. (2020) discusses units that should be used for annotation

of poetic texts. Thus, the authors propose to follow the logical structure of a poetic work, distinguishing between lines, strophes and the text itself (poems).

In our case, we follow the same approach, however, understanding the hymn as a separate text. The markup that is introduced in our corpus

Unit	Number of Units
tokens	33,266
words	27,191
sentences	2,011
hymns	232
strophes	1,149

Table 1: Number of units.

includes the following levels: characteristics of the hymn and characteristics of the strophe (see Table 1 for statistics). The former include id, ancestor, year, category, category 2, rubric, while the latter are represented by id, original, year. Below we will dwell on each of them in more detail.

4.1 Hymn id

The "*hymn id*" is understood as its number in the book, which in the traditional annotation scheme in a corpus corresponds to a text title. In total, we described 232 hymns.

4.2 Hymn year

The parameter means the year of creation of the hymn as a whole, that is, the latest date (in case the hymn contain several parts, then the text is dated to an earlier period, i.e. its earliest part). For some hymns two dates may be given, indicating that the strophes were composed at different times and therefore belong to different periods (1940/1970). The most "productive" (in terms of tokens) period includes the 16th century, 17th century and the first half of the 20th century. The average length of a hymn is 143 words. The longest hymns, containing more than 300 words, were written in different periods: the second half of the 20th century (1959/1972), the 16th century (1537 and 1599), the 17th century (1656) and the 18th century (1771). The shortest hymns date mainly from the 16th century. (1522, 1528-1529, 1531), although short hymns from the 20th century are also found (1947 and 1965). Their length varies from 32 up to 45 words.

4.3 Hymn ancestor

Since, as already indicated, the corpus is conceived as a source of data for studying the genesis of a church hymn (its origins, development and transformation), we paid attention to whether the original hymn exists. The "*hymn_ancestor*" parameter indicates whether the ancestor text exists, and can be "yes" or "no". The majority of hymns (91.4%) have no preceding text.

4.4 Hymn category

The hymn category denotes the century of hymn creation: before the 16th century, 16th century, 17th century, 18th century, 19th century, first or second half of the 20th century.

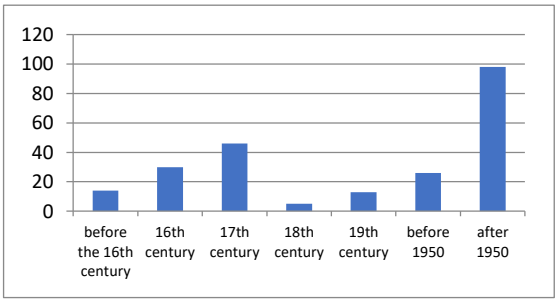

Figure 1: Distribution of hymns (time).

In Fig. 1, bar chart shows the distribution of hymns by time category.

4.5 Hymn category 2

For a more accurate differentiation between hymns by time periods and in order to take into account their musical accompaniment, the texts were divided depending on whether the musical arrangement corresponded to the time of creation of the text (category T+M, namely text and music) or whether the text received a completely different musical arrangement, usually from a later period (category T, namely text only). The texts were marked using appropriate numerical tags (Table 2). Thus, hymn_category_2="4.1" for the above-mentioned hymn 615, means that it belongs to the 18th century and has musical accompaniment corresponding to the time of its creation.

In general, more than half of the hymns (about 60%) have musical accompaniment created during the period when the text was written.

	before the 16th cent.	16th cent.	17th cent.	18th cent.	19th cent.	before 1950	after 1950
	1	2	3	4	5	6	7
T	1.0	2.0	3.0	4.0	5.0	6.0	7.0
T+M	1.1.	2.1	3.1	4.1	5.1	6.1	7.1
	MA	R	B	AE	H	LR	A

Table 2: Musical accompaniment in hymns.

We assume that the division into time categories may also correspond to the correlation of hymn texts by significant cultural and historical periods:

- MA – Middle Ages;
- R – Reformation;
- B – Baroque;
- AE – the Age of Enlightenment;
- H – historicism;
- LR – liturgical Renaissance in Germany;
- A – the era of K. Adenauer and the Federal Republic of Germany.

4.6 Hymn rubric

In this case, the labels were applied from the perspective of pastoral-theological categorization, meaning the liturgical rubric. The parameter is important in the study of liturgical texts. But the division that exists in the hymnbook does not always correspond to the time of the church year, hence we observe a number of discrepancies. For example, in the church year, the period of Lent and most of Holy Week are combined together, while the eve of Easter already refers to Easter time. In Gotteslob (1975), the chants of Lent and the chants of Holy Week are separated. At the same time, the hymns for the eve of Easter are included in the circle of hymns for Holy Week.

Additional expert classification took into account the division of hymns corresponding to liturgical time, but also retained the division in the hymnbook that reflects the traditions of German liturgical practice (Table 3 shows the example).

hymns ID	Gotteslob (1975)	hymns ID	Expert classification
195-211	Karwoche	207-256	Osterzeit
288-311	Vertrauen und Bitte	541-547	Fronleichnam
425-540	Messgesänge	548-568	Jesus Christus

Table 3: An example of hymn rubrics.

In total, 18 rubrics were marked in the hymnbook (marked with the "*Hymn_rubric_gl*" parameter), while the expert identified 14 items (marked with the "*Hymn_rubric_expert*" parameter).

4.7 Strophe id

The parameter marks the strophe number within the hymn. A hymn can have from 1 to 15 strophes. For example, hymn 518, consisting of the maximum number of strophes (15 strophes), dates back to 1962, while hymns 130 and 156, consisting of 14 strophes, date back to 1962 and 1959/1972, respectively.

4.8 Strophe original

A hymn may consist of strophes from different time periods, so for each strophe it is important to indicate whether it is original or not. This logical parameter can take the value "yes" or "no". Most of the strophes (95.7%) are original.

4.9 Strophe year

The parameter indicates the year when the strophe was written. As noted above, in a number of hymns it is the same for all strophes; in other cases, the first strophe is written earlier, while the rest are later. For the top 10, the classification of strophes by year generally repeats the classification of hymns by year, however, the 20th century prevails, because the strophes were written and added to the rest of hymns during this period.

5 Conclusion

This article presents the results of creating a corpus of Catholic hymns, which is the first attempt at developing such a corpus. This material has not been considered in existing text collections, databases, and corpora until now.

The paper described the initial results obtained for metatextual markup of texts. Currently, we plan to take into account such parameters of the verse as meter and rhyme. The next stage will include the morphological annotation of hymns, which will also require additional preparation, because automatic processing of diachronic texts can be tricky and laborious. Future work will include topic modeling of hymns and their clustering in order to identify interrelations between texts.

References

Berliner Repertorium (BR), https://repertorium.sprachen.hu-berlin.de

Deutsch Diachron Digital (DDL), https://www.deutschdiachrondigital.de/

Deutsches Lyrik Korpus (DLK), https://github.com/tnhaider/DLK

Deutsches Textarchiv (DTA), https://www.deutschestextarchiv.de/

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.

John McHardy Sinclair and J. Ball. 1996. EAGLES. Preliminary Recommendations on Text Typology. [online] https://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html

Mette-Marie Møller Svendsen, Nicolai Hartvig Sørensen, and Thomas Troelsgård. 2020. An automatically generated Danish Renaissance Dictionary. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 29–32, Marseille, France. European Language Resources Association.

PO-EMO Corpus, https://github.com/tnhaider/poetry-emotion

Russian National Corpus. http://ruscorpora.ru

Text Encoding Initiative. https://tei-c.org

Thomas Haider and Steffen Eger. 2019. Semantic Change and Emerging Tropes in a Large Corpus of New High German Poetry. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages, 216–222.

Thomas Haider, Steffen Eger, Evgeny Kim, Roman Klinger and Winfried Menninghaus. 2020. PO-EMO: Conceptualization, Annotation, and Modeling of Aesthetic Emotions in German and English Poetry. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1652–1663, Marseille, France. European Language Resources Association.

Thomas Haider. 2021. Metrical Tagging in the Wild: Building and Annotating Poetry Corpora with Rhythmic Features. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3715–3725, Online. Association for Computational Linguistics.