Towards a GoldenHymns Dataset for Studying Diachronic Trends in 19th Century Danish Religious Hymns

Ea Lindhardt Overgaard
School of Communication and Culture
Aarhus University
elt@cc.au.dk

Pascale Feldkamp
Center for Humanities Computing
Aarhus University
pascale.moreira@cc.au.dk

Yuri Bizzoni
Center for Humanities Computing
Aarhus University
yuri.bizzoni@cc.au.dk

Abstract

Religious hymns represent a particularly complex literary domain, due to their specialized registry and context, which remain understudied in computational linguistics, especially in less-resourced languages. We introduce GoldenHymns(S), a novel dataset of Danish historical religious hymns (1798–1873). To allow for a comparison with existing NLP tools' performances, the GoldenHymns(S) dataset is enriched with modernized Danish versions as well as English translations of the hymns. To further the study of sentiment changes in Danish religious texts - a particularly relevant aspect of their development - we also provide verse-level valence annotations by human experts, and we examine the effect of language change and specificity on the performance of contemporary Danish sentiment analysis tools. The dataset is the first resource for evaluating and enhancing the performance of sentiment analysis within the realm of historical religious poetry in the Danish language.

1 Introduction

Historical texts present a challenge for the performance of computational models, especially in under-resourced languages (Schmidt et al., 2021; Zilio et al., 2024). Danish religious hymns represent just such a challenge for computational tools. The hymnal tradition has a central place in Danish culture (Nielsen, 2020), and the official hymn books, which disseminates and curates the hymnal heritage, is the most widely distributed book in the poetry genre in Denmark (Sandstrøm, 2007). Communal singing ("fællessang") remains popular, continually drawing on and revitalizing the hymnal heritage (Baunvig, 2020), especially the production of central hymnists of the late 18th and 19th century. It is a common interpretation that older hymns depict a dualistic view of the world, where a dominant eschatological understanding highlights earthly instances more negatively and religious instances more positively than in newer hymns. A hymn dataset provided with valence scores makes investigations in such polarity shifts possibly both regarding certain religious concepts within the hymns as well as general sentiment structures in the chronology of the hymns. During the 19th century, a historically critical period for Danish society (Glenthøj and Ottosen, 2021) and culture alike (Mortensen et al., 2006),[1] a long line of hymnists and poets took up the endeavor to boost Danish literature and contribute to official church hymn books,[2] which theologians and civil servants wished to "express happy feelings" (Kjærgaard, 2003). Hymnists brought deism and Enlightenment ideals into a new Christian outlook (Baunvig and Nielbo, 2023; Nielsen, 2020), which echoes in the continued publication of official hymn books to this day. This change in mood in 19th century hymns remains relatively unexplored, with studies focusing on individual authors over historical trends (Baunvig, 2023), especially in traditional humanities research (Elbek, 1959).

In this paper, we present two main contributions to the quantitative exploration of historical sentiment changes in religious hymns:

1) We present a new dataset – the Golden-Hymns(S)[3] – of Danish historical religious hymns

---

[1]The period is known as *the Danish Golden Age* for its cultural and political productivity.

[2]Official church hymn books were released in Denmark in 1569.

[3]The "Small" (S) designation reflects its current composition of 65 hymns, with plans for expansion in the future.

with human-annotated valence scores. [4]

2) We evaluate different sentiment analysis (SA) tools on the dataset. To distinguish the effects of language change on the systems' performance, we repeat our experiments on a version translated into modern Danish.

2 Related Works

2.1 Available resources

There are currently three datasets for Danish hymns available. The first is the work of one author, N.F.S. Grundtvig (CGR, 2019), where the hymnal collection is only a subset of the full authorship of Grundtvig. The second dataset indexes hymnal books and bibles related to the reformation era (DSLDK, 2021), spanning from 1529 to 1569 – where only the latter collection, from 1569, is an authorized hymnal book.[5] A third dataset consists of the hymns from the most recent authorized Danish hymnal book of 2002 (By-, Land- og Kirkeministeriet and Vajsenhus, 2002). Beyond the most modern one, the existing datasets treat one authorship or specific period. As such, there is to the best of our knowledge no dataset that facilitates the study of diachronic change in the genre.

The second (Reformation) dataset does enable study of the development of a (limited) period, as well as comparative analysis of historical characteristics and registry of religious Danish in the genre. Yet, the hymns included in the dataset feature comprehensive orthographic deviations (compared to modern Danish), posing a challenge to human annotators with its 16th century Danish, possibly resulting in low levels of agreement between annotators.

With the GoldenHymns(S) dataset, we thus supplement available Danish resources, facilitating both diachronic study and supplying texts that have been modernized and annotated by scholars experienced with the linguistic registry.

2.2 Sentiment Analysis of historical texts

While the study of sentiment has come to be a central approach in computational literary studies (Rebora, 2023), its applications to Danish literature – especially historical – are relatively rare. Though significantly less resourced than English, there is

---

	No.	Vs	Words	$\bar{x}$ Vs	Period
Hymns	65	1,914	10,303	32.9	1798-1873

Table 1: Presenting the dataset: The total number of verses (Vs) and words, mean ($\bar{x}$) number of verses per hymn, and timeframe.

no obvious issue halting the use of SA tools in the Danish language and literature: Danish dictionary-based tools show comparable performance (Schneidermann and Pedersen, 2022), and a tool like Sentida has been validated against human scores of text (chunks) across domains, showing a robust performance for fiction (Lauridsen et al., 2019). Still, there is no comprehensive Danish SA benchmark, and the performance is generally evaluated on modern Danish across (few) domains. Assessing the performance of SA models on historical Danish and Norwegian literary texts, Allaith et al. (2023) found that multilingual transformer models outperformed models trained on modern Danish as well as classifiers based on Danish lexical resources. Schmidt et al. (2021) similarly found that transformers did best on historical German drama. Testing dictionary-based methods on historical German plays, Schmidt and Burghardt (2018) found that dictionaries did well when extending their lexica with historical variants. They also found an issue with low levels of agreement between annotators who were not used to the historical register – foregrounding the importance of using annotators experienced with the register of the period and domain. Toft (2023) addressed how modernizing historical Danish hymns improved the performance of Danish NLP tools achieving significant improvements in the accuracy of automatic annotating of POS-tags using models like DaCy and SpaCy, underscoring the importance of either normalizing historical language or adapting NLP techniques when working with historical texts.

3 Dataset

The dataset consists of 65 Danish hymns collected at random from three different official hymnal books from the years 1798 (n=35), 1857 (n=17), and 1873 (n=13) (psa, 1798, 1857, 1873). Popular hymns from earlier periods might be included in these collections, but following Danish tradition, hymns in the hymnals were edited from their authentic version into a version that fitted the church at the time. Each hymn is then edition-specific and

represents the associated hymnal. The hymns are characterized by their metrical structure (verses, rhymes, etc.), their poetic language style (exclamations, figurative language), and their archaic and formal language - e.g., the use of Latinized "est" for "is" ("er"). Each hymn is coupled with its modernized version and English translation (which maintain the original verse style and syntactical structure), as well as a sentiment score per verse.

Hymns are challenging for SA due to their register. Their poetic and figurative language, often embodies subtle emotional tones, as well as the cultural and religious contexts they refer to (Skovsted et al.; Nielsen, 2020). Due to the poetic style, the sentiment analysis is based on dividing the hymns into verses (Table 1). Our unit of analysis was verses since the fundamental unit of poetry is the verse, rather than the sentence. A syntactically sound sentence might thus not be present in every verse, which further challenges sentimental interpretations.

By including both original and modernized versions of the hymns in the dataset we allow observations on how modernizing the hymns affects Danish sentiment analysis. Furthermore, by including validated English translations, we provide accessibility for English-speaking researchers and cross-linguistic comparability (e.g. comparing English SA model performance). Examples of verses from the dataset is shown in Table 2, with examples of the original verse, its modernization and its English translation.

Sentence	Score
J Mistvivl, Angest, Smerte (M) I mistvivl, angst og smerte (En) In doubt, anxiety, and pain	2.0
Ungdomsliv i Morgenrøden (M) Ungdomsliv i morgenrøden (En) Youthful life in the morning's red glow	7.0

Table 2: Example of a positive and negative sentence (original, modernized and English translation) with the human mean score.

4 Methods

We provide an overview of how the dataset was supplemented and annotated for valence (by human annotators and automatic systems). We then show a use-case of the valence annotation, comparing human to automatic scores – both overall for original and modernized versions of the hymns and for each of the hymn book collections separately – to examine systems' performance and how it may vary throughout the period covered by our data.

4.1 Modernization and translation

Beyond the original hymn texts, the dataset includes modernized versions and English translations of each verse following the original verse and syntactical structure. The verses were modernized by two scholars, who prompted ChatGPT 3.5,[6] and subsequently validated each output verse manually against the original to ensure that spelling and vocabulary were updated, keeping semantic and syntactical changes to a minimum. English translations were created by using Google Translate, and then again manually revised and validated by two bilingual experts.

4.2 Human Sentiment Annotation

Danish language and literature scholars (n=2)[7] read and scored all 1,914 verses – defined by line breaks – on a 0 to 10 valence scale: 0 signifying the lowest, and 10 the highest valence (for example sentences see Table 2). Here, valence was intended as the sentiment expressed by the verse. Annotators were instructed to try reporting on the sentiments embedded in the verse, i.e., to think about the valence of each verse individually, not overthinking context.

We report a relatively high inter-rater correlation, with a Spearman's r between their scores of 0.726 – high considering the fragmentary nature of the text rated (verses, not sentences) and considering that humans rarely have an agreement higher than 80% for tasks like positive/neutral/negative tagging (Wilson et al., 2005) or 0.80 Krippendorff's α for continuous scale polarity annotation of non-fiction texts (Batanović et al., 2020).[8]

4.3 Automatic Annotation

We used several models on Danish sentiment analysis, both transformer- and dictionary-based (the latter of which are usually also rule-based), to score the verses in the hymns for valence. Dictionary-based methods remain popular due to their transparency and versatility, and appear to perform well

	Alex.inst.	Senda	RoBERTa	Asent	Afinn	Sentida
Hymns original	0.39	0.32	0.39	0.40	0.39	**0.49**
Hymns modernized	0.42	0.35	0.46	0.41	0.40	**0.53**

Table 3: Sentiment analysis of hymns: Spearman correlation between scores on the **original** (above) and **modernized lines** (below) to the human mean scores (annnotated on the original lines). Transfomer-based systems are on the left, and dictionary and rule-based systems are on the right. For the correlations, all pvalues are <0.01.

	Alex.inst.	Senda	RoBERTa	Asent	Afinn	Sentida
1798	0.29 (0.36)	0.30 (0.36)	0.36 (0.40)	0.35 (0.33)	0.36 (0.34)	**0.43 (0.44)**
1857	0.39 (0.36)	0.36 (0.38)	0.43 (0.51)	0.41 (0.42)	0.38 (0.39)	**0.49 (0.53)**
1873	0.43 (0.48)	0.30 (0.33)	0.37 (0.46)	0.40 (0.44)	0.40 (0.44)	**0.51 (0.56)**

Table 4: Sentiment analysis over time: Spearman correlation between scores on the **original** and **modernized lines** (in parentheses) to the human mean scores for each hymn-collection individually (1798, 1857, and 1873). Transfomer-based systems on the left, dictionary and rule-based systems on the right. For the correlations, all pvalues are <0.01.

on literary texts ([Bizzoni and Feldkamp, 2023]). We test sentiment dictionaries that have had a wide application in Danish. Our chosen dictionary-based tools were:

Afinn: A valence dictionary without rules, created from Twitter data and various open sources.[9] The dictionary includes many inflections of the same lemma. Valence scores range from -5 to +5.

Sentida: A rule-based system inspired by the English VADER, considering negations, adverb modifiers, and more.[10] Sentida integrates the Afinn dictionary with the 10,000 most frequent Danish lemmas, which were manually annotated by the authors ([Lauridsen et al., 2019]). It relies on stemming to find matching dictionary entries during inference. Valence scores range from -5 to +5.

Asent: A rule-based system that is part of the DaCy suite (asent_da_v1), a comprehensive NLP toolkit for Danish.[11] It uses the Afinn dictionary by default and adds rules for handling negations, modifiers, intensifiers, etc. Scores range from -1 to +1.

Moreover, we use more recent Transformer-based models, which are becoming popular and show potential applied to literary texts ([Elkins, 2022]) and historical literary texts ([Allaith et al., 2023]). We chose to use two off-the-shelf models currently developed for Danish SA, as well as one widely used multilingual model, RoBERTa xlm, which has shown a good performance on literary prose ([Biz-zoni and Feldkamp, 2023]):[12]

Senda:[13] was specifically created for Danish. It is based on the Roberta architecture and pretrained on an extensive collection of Danish texts.

Alexandra Institute sentiment base:[14] represents another Danish-oriented transformer model fine-tuned for sentiment analysis tasks. This model is provided by the Alexandra Institute.

RoBERTa xlm multilingual base sentiment:[15] uses cross-lingual training techniques, designed to enhance its capacity for understanding and processing multiple languages by transferring knowledge. This method allows the model to apply skills from one language to another, which can improve its generalization in sentiment analysis. However, this may limit its effectiveness with language-specific nuances, particularly in specialized domains.

We opted to exclude GPT models or new generation LLMs at this stage. Generative models like GPT suffer from increased opacity even with respect to traditional transformers and, in the case of the largest models, are trained on unknown data. Most importantly, GPTs' generative nature makes any application dependent on prompting, which introduces a level of variability and inconsistency that would complicate our results. While this was

---

[9]https://github.com/fnielsen/afinn
[10]https://github.com/Guscode/Sentida
[11]https://centre-for-humanities-computing.github.io/DaCy

[12]We maintained all presets as the default when applying these models, so that the hyperparameters are as specified in the documentation of the individual model (see the model hyperlinks).
[13]https://huggingface.co/larskjeldgaard/senda
[14]https://huggingface.co/alexandrainst/da-sentiment-base
[15]https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment

not a problem in the text modernization phase, as it made sense to review and correct the result, we opted to stick to tools specifically trained to return an SA output from a text, leaving the introduction of generative models to a later phase.

5   Results

We report the performance (Spearman's *rho*) of system sentiment scores compared to human scores in Table 3. It appears that dictionary-based tools (Afinn, Sentida and Asent) perform comparable to or – in the case of Sentida – better than transformer-based methods. Moreover, all methods improve when applied to modernized texts, against the original text, with the biggest improvements observed for Sentida (+0.04) and RoBERTa (+0.07). Across the collections, published in 1798, 1857, and 1873, models generally improve over time. Notably this trend is weakened – though not gone – when applied to the modernized versions of the text (Table 4). Irrespective of the year of publication, Sentida continues to perform the best in this genre.

Among transformer-based models, the Alexandra Institute and RoBERTa xlm models appear to perform best, and the consistency of their performance over time appears only slightly worse than dictionary-based models (e.g., a 0.12 point difference for the Alexandra Institute model, and an 0.8 point difference for Sentida when contrasting coefficients of the earliest and the latest collection).

6   Discussion and Conclusions

Contrary to previous findings on modern narrative (Bizzoni and Feldkamp, 2023) and historical drama and narrative texts (Allaith et al., 2023; Schmidt et al., 2021), where multilingual transformer-based models appeared to perform best, we find that dictionary-based methods outperform transformer-based models in this domain. Transformer-based models may depend more heavily on the presence of syntactically sound phrases (which are not always extant in hymn verses), and be impaired by a verse-level tokenization of the hymns, worsening their performance. It is interesting to note that systems that put particular emphasis on the valence of individual words appear to perform better in contexts – such as a single poetic verse – where semantic blow of single words is more central. An intriguing observation of this study is the overall improvement of models through time, which also holds true when applying the models to modernized verses. It is possible that hymns evoke sentiment differently through time, with hymnists changing their affective strategies, so that factors like the level of language concreteness, which has been shown to elude SA models and impact perceived sentiment in literary texts (Bizzoni and Feldkamp, 2024), might change through time.

Studies on sentiment analysis (SA) within the domain of historical hymns can contribute to the broader field of SA in literary genres. Hymns, with their consistent and limited themes, use of topics and metaphors, serve as a good starting point for exploring SA in poetry in general. This consistency provides a controlled environment to refine and test SA methodologies, as the genre is quite accessible to machine processing, which can then be applied to more complex and varied literary forms. Insights gained from SA in hymn literature not only enhance our understanding of emotional expression in religious and historical texts but also offer valuable methodologies that can be adapted for other literary genres. For instance, the success of dictionary-based models in outperforming transformer-based models within the context of hymn analysis suggests that traditional, lexicon-driven approaches may have advantages in certain types of literary analysis such as poetis genres. This could be due to the repetitive and formulaic nature of hymns, which might be better captured by dictionary-based models. By starting with hymns, researchers can develop robust techniques that address the unique challenges of poetic texts, such as figurative language and complex emotional expressions. These techniques can then be extended to analyze sentiment in a wide range of literary genres, from classical poetry to modern prose, thus enriching the field of SA in literature as a whole.

Overall, insights from the current study show ample reason to work towards expanding this dataset in the future, both in terms of size and temporal range. Future work might test the improvement of SA tools' performance when supplying lexica of historical language variants, and examine the development in hymnal sentiment across time generally or in connection with certain concepts. We also plan to work on larger datasets and to fine-tune models on this kind of sentiment annotations. Finally, an undoubtedly interesting next step would be that of testing the behaviour of generative large language models on the task, both in terms of similarity with human judgments and in terms of prompt-dependent variability.

References

1798. *Evangelisk-kristelig Psalmebog til Brug ved Kirke- og Huus-Andagt.* Kongl. Vaisenhus, Kiøbenhavn.

1857. *Nyt Tillæg til evangelisk-christelig Psalmebog.* Kongl. Vaisenhus, Kiøbenhavn.

1873. *Tillæg til Psalmebog for Kirke- og Huus-Andagt.* Kongl. Vajsenhuses Forlag, Kjøbenhavn.

Ali Allaith, Kirstine Degn, Alexander Conroy, Bolette Pedersen, Jens Bjerring-Hansen, and Daniel Hershcovich. 2023. Sentiment Classification of Historical Danish and Norwegian Literary Texts. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 324–334, Tórshavn, Faroe Islands. University of Tartu Library.

Vuk Batanović, Miloš Cvetanović, and Boško Nikolić. 2020. A versatile framework for resource-limited sentiment articulation, annotation, and analysis of short texts. *PLoS ONE*, 15(11):e0242050.

Katrine F. Baunvig and Kristoffer Nielbo. 2023. Benign Structures. The Worldview of Danish National Poet, Pastor, and Politician N.F.S. Grundtvig (1783-1872). *Digital Humanities in the Nordic and Baltic Countries Publications*, 5(1):1–10. Number: 1.

Katrine Frøkjær Baunvig. 2020. Forestillede fællesskabers virtuelle sangritualer: Forskningsprojekt vil kaste lys over den kulturelle betydning af den virtuelle fællessang under corona-tiden. *Tidsskriftet SANG*, 1(1):40–45. Number: 1.

Katrine Frøkjær Baunvig. 2023. A Computational Future? Distant Reading in the Historical Study of Religion. In *Stepping Back and Looking Ahead: Twelve Years of Studying Religious Contact at the Käte Hamburger Kolleg Bochum*, pages 325–352. Brill. Section: Stepping Back and Looking Ahead: Twelve Years of Studying Religious Contact at the Käte Hamburger Kolleg Bochum.

Yuri Bizzoni and Pascale Feldkamp. 2023. Comparing transformer and dictionary-based sentiment models for literary texts: Hemingway as a case-study. In *Proceedings of the 3rd International Workshop on Natural Language Processing for Digital Humanities*, pages 219–226, Tokyo, Japan. Association for Computational Linguistics.

Yuri Bizzoni and Pascale Feldkamp. 2024. Below the sea (with the sharks): Probing textual features of implicit sentiment in a literary case-study. In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 54–61, Malta. Association for Computational Linguistics.

By-, Land- og Kirkeministeriet and Det Kgl. Vajsenhus. 2002. Den danske salmebog online.

CGR. 2019. Salmer Dataset. Centre for Grundtvig Research (CGR), GitHub repository.

DSLDK. 2021. Salmer Dataset. Society for Danish Language and Literature (DSLDK), GitHub repository.

Jørgen Elbek. 1959. Grundtvig og de latinske salmer. *Grundtvig-Studier*, 12(1):7–64.

Katherine Elkins. 2022. *The Shapes of Stories: Sentiment Analysis for Narrative*. Cambridge University Press.

Rasmus Glenthøj and Morten Nordhagen Ottosen. 2021. *Union eller undergang: kampen for et forenet Skandinavien*. Gads Forlag.

Jørgen Kjærgaard. 2003. *Salmehåndbog*. Det Kgl. Vajsenhus' Forlag.

Gustav Aarup Lauridsen, Jacob Aarup Dalsgaard, and Lars Kjartan Bacher Svendsen. 2019. SENTIDA: A New Tool for Sentiment Analysis in Danish. *Journal of Language Works - Sprogvidenskabeligt Studentertidsskrift*, 4(1):38–53. Number: 1.

Klaus P. Mortensen, May Schack, and Annemette Sørensen, editors. 2006. *Dansk litteraturs historie*. Gyldendal.

Marita A. Nielsen. 2020. Salmesprog. In *Dansk Sproghistorie Bind 4. Sprog i brug.* Aarhus University Press and Society for Danish Language and Literature (DSLDK).

Simone Rebora. 2023. Sentiment Analysis in Literary Studies. A Critical Survey. *Digital Humanities Quarterly*, 17(2).

Bjarne Sandstrøm. 2007. Salmen - fra kampsang til lovprisning. In V. A. Pedersen, M. Schack, and K. P. Mortensen, editors, *Dansk Litteraturs Historie 1100-1800*. Gyldendal.

Thomas Schmidt and Manuel Burghardt. 2018. An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 139–149, Santa Fe, New Mexico. Association for Computational Linguistics.

Thomas Schmidt, Katrin Dennerlein, and Christian Wolff. 2021. Using Deep Learning for Emotion Analysis of 18th and 19th Century German Plays. In Manuel Burghardt, Lisa Dieckmann, Timo Steyer, Peer Trilcke, Niels-Oliver Walkowski, Joëlle Weis, and Ulrike Wuttke, editors, *Fabrikation von Erkenntnis: Experimente in den Digital Humanities*, pages 1–32. Melusina Press.

Nina Schneidermann and Bolette Pedersen. 2022. Evaluating a New Danish Sentiment Resource: the Danish Sentiment Lexicon, DSL. In *Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data*, pages 19–24, Marseille, France. European Language Resources Association.

Morten Skovsted, Mads Djernes, Kirsten Nielsen, Martin Horsntrup, and Hanne J. Jakobsen. Hvad gør en ny salme til en god salme? *Salmedatabasen*.

Ea Lindhardt Toft. 2023. 1500-talssalmer og danske sprogmodeller. *19. Møde om Udforskningen af Dansk Sprog*.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Leonardo Zilio, Rafaela Radünz Lazzari, and Maria Jose Bocorny Finatto. 2024. NLP for historical Portuguese: Analysing 18th-century medical texts. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 76–85, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.