Universiteit
Leiden
The Netherlands

Popular song topics in the Dutch Republic: a data-driven study into topical fluctuations in the Dutch song database (1550-1750)

Lassche, A.W.

Citation
Lassche, A. W. (2022). Popular song topics in the Dutch Republic: a data-driven study into topical fluctuations in the Dutch song database (1550-1750). *Early Modern Low Countries*, *6*(2), 253-277. doi:10.51750/emlc10908

Version:	Publisher's Version
License:	Creative Commons CC BY-NC 4.0 license
Downloaded from:	https://hdl.handle.net/1887/3512131

Note: To cite this publication please use the final published version (if applicable).

Popular Song Topics in the Dutch Republic: A Data-Driven Study into Topical Fluctuations in the Dutch Song Database (1550-1750)

Alie Lassche

Alie Lassche is a PhD candidate in Dutch History at Leiden University within the project 'Chronicling Novelty. New knowledge in the Netherlands, 1500-1850'. She investigates the changing mediascape of Dutch chroniclers, using both qualitative and quantitative methods to analyse a corpus of about three hundred early modern Dutch chronicles. In 2019, she was awarded her research master's degree in Dutch literature and culture at Utrecht University with distinction. In her research, she applies computational methods to historical text corpora, in order to investigate cultural changes over time.

Abstract

This article investigates popular topics and topical fluctuations in a diachronic corpus of 43,772 Dutch songs, all written between 1550 and 1750, contained within the Dutch Song Database. Computational methods such as topic modelling are used to analyse the relationship between topical changes and cultural-historical developments. Two cultural trends are used as case studies: the role of Petrarchism, and the articulation of a patriotic identity in early modern Dutch song culture. Furthermore, this data-driven approach reveals how subcategories can be defined within the existing but incomplete genre classification in the song collection. The results obtained contribute to a better understanding of the richness of the Dutch Song Database, and will facilitate the use of the song collection by future users.

Keywords: song culture, topic modelling, topical fluctuations, text mining, digital humanities

DOI 10.51750/emlc.10908 - URL: http://www.emlc-journal.org
Publisher: Stichting EMLC
Copyright: The Author(s). This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Popular Song Topics in the Dutch Republic: A Data-Driven Study into Topical Fluctuations in the Dutch Song Database (1550-1750)

Alie Lassche

During the early modern period, the tradition of the monodic song in the vernacular remained extremely strong in the Dutch Republic. Tens of thousands of Dutch songs were written, published, read, and, obviously, sung. Young and old, rich and poor alike, were engaged in song culture, which was, according to Louis Peter Grijp, as typically Dutch as the culture of painting.[1] According to Natascha Veldhorst, singing was second nature to the Dutch: everyone sang anywhere and at any time.[2] Traces of this practice can be found not only in contemporary paintings, but also in archives and libraries, where hundreds of early modern Dutch songbooks are still kept. The singing tradition flourished especially in the seventeenth century. Previous studies have already shown that singing was a social business in the early modern Low Countries; people usually sang together and children were taught to sing in school. Singing and music lessons had been a common part of the curriculum from the sixteenth century, as *musica* was considered one of the *artes liberales*. Even those who could not read music were able to take part: the use of the contrafactum (writing new lyrics on existing melodies) made it possible for such singers to learn new songs anyway.[3]

But what did people sing about during early modern times? And what changes in song topics can we observe over a long period of time? For decades, literary scholars have published studies on historical songs, their topics, the public, and their function in society. Eddy Grootes, for instance, has analysed the rise in popularity of certain songbooks for youngsters in the first quarter of the seventeenth century.[4] Els Stronks has researched early

1 Grijp, *Het Nederlandse lied*, 29. This article is based on the thesis I wrote to complete the research master's degree in Dutch literature and culture at Utrecht University in July 2019, supervised by Els Stronks and Folgert Karsdorp. I would like to thank Joris Veerbeek for being my Python lifeline, and Lucas van der Deijl, Milan van Lange, Judith Pollmann, Kim Smeenk, the editorial board of EMLC, and the anonymous reviewers for reading and commenting on earlier versions of this article. The code that was used in this study can be found at https://github.com/awlassche/dutch-historical-song-topics (accessed on 18 October 2022).

2 Veldhorst, *Zingend door het leven*, 12.

3 Stronks, '"Dees kennisse zuldy te kope vinnen"', 152.

4 Grootes, 'Het jeugdig publiek'.

modern songs written by and for youngsters, demonstrating how songs played a role in the identity formation of the early modern Dutch youth.[5] Furthermore, she has argued how songs were used to transmit knowledge.[6] The most famous Dutch song from the early modern period was chosen to be the national anthem of the Netherlands in 1932. This song – the *Wilhelmus* – is without doubt the most researched song in Dutch history, not least because the identity of its composer was shrouded in mystery for many years. This mystery was finally solved in 2017 through stylometrics, in the notable study *Van wie is het Wilhelmus?* ('Who wrote the *Wilhelmus*?').[7]

Although many authors have studied the topics of early modern songs, their research was always based on case studies. Using such methodology does not enable the researcher to search for patterns that occur over a longer period of time. Quantitative studies on early modern songs are in fact scarce, and this scarcity is closely related to the supposed lack of data, a problem already observed by Veldhorst. In her work on Dutch early modern songbooks, she noted that it was hard to come up with specific numbers regarding early modern song production. In the first place, there is no overview of those early modern songbooks currently held in European libraries. Secondly, one has to keep in mind that songbooks were heavily modified, expanded, and revised over the course of time. Thirdly, the small survival chances of the songbook further complicate any attempt at estimating the number of songs that were printed at the time.[8] And if there are two things you need in order to perform quantitative research, it is numbers and data.

While we may lack comprehensive data sets on early modern Dutch songs, that does not mean we have no representative data at all. In the early nineties, the musicologist Louis Grijp undertook an impressive and successful attempt to map Dutch early modern song culture. In his 1991 dissertation he reconstructed the melodies of songs by finding their sources and analysing their reception. One of the results of his extensive work was the birth of the so-called *voetenbank*, an electronic database containing data on melodies, verses, and stanzas of (at the time) 5,700 Dutch historical songs.[9] Almost thirty years after Grijp's dissertation, his database has developed into an extensive Dutch Song Database, containing metadata of more than 175,000 Dutch songs. Furthermore, in the last couple of decades the techniques for digital exploration of a corpus have become more advanced. This article shows how making full use of these new techniques can allow for meaningful quantitative research into early modern Dutch songs and how their dominant topics changed over time. I have established what topics dominated early modern Dutch songs by performing topic modelling – a way to identify the topics that occur in a collection of documents – on a corpus of 43,772 Dutch songs from the period 1550-1750.

The aims of this article are manifold. Firstly, it aims at contributing to a better understanding of the richness of the Dutch Song Database. At the point of writing, it lacks comprehensive metadata – an example being information regarding song genre, which

5 Stronks, 'Identiteiten van adolescenten', 225.
6 Stronks, '"Dees kennisse zuldy te kope vinnen"'.
7 Kestemont et al., *Van wie is het Wilhelmus*?
8 Veldhorst, 'Pharmacy', 225-227.
9 Grijp, *Het Nederlandse lied*, 13.

makes the search for songs of a specific genre a frustrating task. The results published in this article may therefore serve to increase the database's usability for future researchers. Secondly, through two case studies I will analyse the dynamics which can be observed in the fluctuation of dominant topics over time, and suggest how these changes in topical distributions can be related to the cultural-historical context in which they occur. Finally, I wish to demonstrate how this methodology can contribute to further research into early modern Dutch songs, and, indeed, into other textual genres.

Although I am, in this study, using methods never before applied to early modern songs, I also rely heavily on previous, qualitative, research into early modern Dutch song culture. The two cases discussed in more detail are well known from existing studies: the role of Petrarchism in early modern Dutch songs, and the way in which songs were used to articulate a patriotic identity – with the *Wilhelmus* being perhaps the classic example. To explore how these two cases can be related to the results of a quantitative approach to researching Dutch song culture, I use a sample of the Dutch Song Database, containing songs from 1550-1750. The article begins by offering a historical background of the two cases. The second section will discuss the corpus and its characteristics in more detail, the third section the methods I have employed, and the two subsequent sections contain my results. The article concludes with an evaluation of how the combination of qualitative and quantitative methods can enlarge our knowledge of early modern song culture.

Petrarchism and Patriotism in Dutch Songs

The starting point of my first case study is that the poetry of early modern Dutch poets became increasingly embedded in a Renaissance and classical tradition. Inspired by the manner in which Italian and French writers adapted the traditions of Greek and Latin poetry, Dutch poets also appropriated their topics and genres. An important example of this embedding in the Renaissance tradition was the practice of the Petrarchan genre. The Italian Renaissance poet Francesco Petrarca – or Petrarch, as he is known in English – was a fourteenth-century Italian poet, whose fame extended beyond his country's own borders already during his lifetime.[10] One of the reasons for Petrarch's fame was the popularity of his love poetry. His love sonnets were distinguished by the worship of a charming but unreachable woman, which meant that love always causes both happiness and pain in his poems. The physical beauty of the beloved is often described using natural metaphors and mythological references. The lover is often characterized as someone with a split personality, who struggles with suicidal tendencies.[11]

In his study on seventeenth-century songbooks for youngsters, Grootes argues that one would not expect Petrarchism to become popular in the Low Countries, since the trope of the unreachable woman was paradoxical to the actual surplus of women in Amsterdam and the province of Holland at the time: 'At first sight, there seems to be no ground for the humbly pleading of the lovers, nor for the presentation of the songbooks

10 Gelderblom, 'Investing in Your Relationship', 131.
11 Van Bork et al., *Algemeen Letterkundig Lexicon*, 'petrarkisme'.

as instruments to make the girls a little more obliging.'[12] Even so, Grootes suggests that the ideology of Petrarchism, with its emphasis on restraint and the lack of sexual boldness, probably fitted perfectly with the marriage policy advocated by the parents.[13] He also cites the literary scholar August Keersmaekers, who has argued that the songbook *Den Nieuwen Lust-hof* (1602) could be interpreted as part of 'a culture of piety propagated and imposed from above', in opposition to the older songbooks.[14] The quantitative methods used in this study are ideally suited to explore whether the popularity of the classic genres, and the Petrarchan genre in particular, can be found in the current research corpus. Furthermore, qualitative research has not paid attention to the moment of decline in popularity of these genres, something which the methodology used in this study might also shed light.

Another claim that will be tested in this study is how songs were used to articulate a feeling of patriotism. As Cornelis van der Haven notes, a long tradition of aggressive patriotic rhetoric can be found in Dutch literature, a tradition that is strongly related to the Dutch liberation myth of the struggle against Habsburg Spain that would be constructed in later centuries.[15] In times of war or political crisis, feelings of patriotism and unity are aroused and propagated by, among others, poems and songs.[16] A social activity, singing together created a feeling of belonging to a particular group, and contributed to the formation and maintenance of a social identity and an imagined community.[17] Regarding the early modern Netherlands, Dieuwke van der Poel, Louis Grijp, and Wim van Anrooij have distinguished seven singing groups. The 'political group' is most relevant in this context, which they describe as a group '(often with a specific religious affiliation) such as the anti-Catholic and anti-Spanish *Geuzen* ('Beggars')'.[18] Their protest songs – known as beggar songs – became very popular in the late sixteenth century. Initially they were published and disseminated via broadsheets, but before long they were bundled together and published in a songbook called *Nieu Geusen Liedtboexcken* (The New Beggars' Songbook). The oldest edition of this song book dates from 1577, and several reprints were published in subsequent years. The songs were intended as inflammatory news announcements that would encourage revolt.[19]

Songbooks were the most obvious way to articulate an identity, in other words, to express one's belonging to a group. Often group identity was already prominent in the title, for example in the *Nieu Geusen liedtboecxken.*[20] Furthermore, identity was also articulated through song lyrics – which is something I will explore in this article. Although the *Wilhelmus* is the most well-known of all beggar songs, and one which retained both its political overtone and signalling function long after the 1648 Peace of Münster, there were

12 Grootes, 'Het jeugdig publiek', 84.
13 Grootes, 'Het jeugdig publiek', 85.
14 Keersmaekers, *Wandelend in Den Nieuwen Lust-Hof*, 119.
15 Van der Haven, 'Patriotism and Bellicism', 56.
16 Jensen (ed.), *The Roots of Nationalism*, 18.
17 Van der Poel, Grijp, and Van Anrooij, *Identity, Intertextuality, and Performance*, 4.
18 Van der Poel, Grijp, and Van Anrooij, *Identity, Intertextuality, and Performance*, 5.
19 Veldhorst, 'Pharmacy', 233.
20 Van der Poel, Grijp, and Van Anrooij, *Identity, Intertextuality, and Performance*, 8.

many other songs.[21] While their popularity probably did not reach the height of the current Dutch national anthem, their tropes of nation and identity likely resonated with the inhabitants of the Dutch Republic. That being said, there were also songs expressing love for the fatherland without specific political orientation, in which the whole Dutch nation was addressed as one political group, especially in times of war. During the seventeenth century, patriotic ideas became more widespread, a movement which was not only down to the beggars and their repertoire. Orthodox Protestant ministers also started to elaborate on the idea that God, the Netherlands, and the House of Orange were inextricably connected.[22]

The Dutch Song Database

To test whether these trends can be found in a larger research corpus, I have relied on the Dutch Song Database, which contains the metadata of approximately 175,000 songs in the Dutch language, from the Middle Ages up to the twentieth century. The database was initiated in the early 1990s by Louis Grijp, who continued to lead its development until 2015. During these years, many research and documentation projects have been carried out, and so an enormous amount of high-quality data has been collected. Nowadays, the Dutch Song Database offers a rich cross-section of Dutch song culture throughout history.[23] It contains several genres, including love songs, satirical songs, psalms and other religious songs, folksongs, and children's songs. For every song, the source for the text and/or the melody are indicated. Other available metadata include the author's name (if known), the first line, the number of stanzas, genre, melody, stanza form, and number of verses. In addition, the results of the project Dutch Songs On Line became available in 2014.[24] The project has made 53,351 full song texts accessible online. Of these texts, 29,590 have had both their lyrics and metadata encoded with TEI compliant XML, which provides the publication date, geographical location, melody, and genre. For the purposes of this study, I have limited this sample to songs that were printed and published between 1550 and 1750, which resulted in a corpus of 22,297 songs. This number does not yet include the reprints and appearances of these songs in other songbooks. Since I wish to examine which song topics were more prevalent than others, it is necessary to include these data as well. Although the number of different lyrics will remain the same, the distribution of the songs can change. Suppose, for example, that songs on politics and history, which are not very prominent in the corpus based on the data in tab. 1, are reprinted so often that their number increases massively, while songs on love are rarely reprinted. This would influence the distribution of topics in such a way that political and historical topics are more dominant than love topics. Expanding the corpus with

21 De Bruin, 'Het Wilhelmus tijdens de Republiek', 24.

22 De Bruin, 'Het Wilhelmus tijdens de Republiek', 38.

23 Van Kranenburg, De Bruin, and Volk, 'Documenting a Song Culture'.

24 This project involved cooperation between the DSD and the Digital Library of Dutch Literature (DBNL), and was funded through a NWO Medium Investments grant.

Tab. 1 Number of songs in the Dutch Song Database by category and average word length. N = 22,297.

Category	Number of songs	Average number of words per song
None	7017	300
Religion	6914	323
Love and sex	3261	219
Seasons and annual events	1033	296
Formal genres	818	334
Amusement	722	251
Emotions	694	322
Narratives	479	363
Cycle of life	473	223
Politics and history	300	478
Groups	187	223
Children	148	221
Occasions	100	277
Theatre	90	207
Work	59	360
Miscellaneous	2	91

Source: Dutch Song Database.

reprints resulted in a corpus of 43,772 songs.[25] Their distribution over time is visualized in fig. 1.

Tab. 1 gives an overview of the distribution of songs over the various categories, as labelled in the Dutch Song Database.[26] The majority of the songs fall into the category 'religion', followed at a distance by 'love and sex' and 'seasons and annual events'. One might ask why a study on topics in song lyrics is useful, given that data on categories already exist, but a closer look at tab. 1 shows the necessity of such a study. Not only does the categorical division seem rather arbitrary, there are also more than seven thousand songs that have not been assigned to a specific category at all. As such, the available data do not give an accurate overview of the distribution of songs over genres. Moreover, the wide-ranging categories do not reveal how subdivisions within a category are distributed. Are all religious songs on the same aspects of religion? Does 'love and sex' mean that some songs are on love, and others on sex? There is an important difference between a genre and a topic, since a genre can comprise many different topics. A song within the genre 'love and sex' might be a tragic love song, but it could also be a happy, uplifting love song. A quantitative analysis of the songs, using topic modelling, can therefore offer useful information: it offers a more fine-grained classification of the different genres, and will allow the topical fluctuations over time to be made visible.

25 Each first print of a song has a *recordid*, which is the id that is mentioned in the XML file of a song. A reprint of a song does not contain a *recordid*, but a *herdrukid* (*reprintid*) instead. Furthermore, each unique song has an *incnormid*. This means that a couple of songs with different *recordids* and *herdrukids* share the same *incnormid*. Songs with the same *incnormid* and/or *recordid* as the ones already in the corpus were added.

26 In recent decades, more than 100 people (researchers, documentalists, interns, and volunteers) have added metadata to the Dutch Song Database. This was mainly done within the context of a project, at the Meertens Institute (which still hosts the database) and at other institutions such as the University of Antwerp.

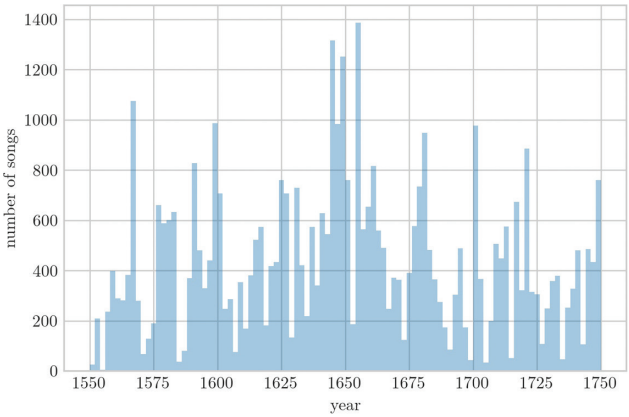

Fig. 1 Annual distribution of songs in the Dutch Song Database, 1550-1750. N = 43,722.

Tab. 1 also provides information about the average song length within each genre. It shows that the average length of a song belonging to the genre 'religion' is about one hundred words longer than a song belonging to the category 'love and sex'. The most outstanding outlier is the genre 'politics and history', whose songs contain, on average, 478 words. The graphs in fig. 2 give a more detailed insight in the distribution of songs within a genre. For the twelve most popular genres, the number of words per song is plotted against their frequency.[27] In general, all genres show a similar distribution: most songs contain between one hundred and three hundred words, with few songs boasting very long word counts. While the genres 'religion', 'love and sex', and the songs categorized with 'none' show a similar distribution, the word length within the genre 'politics and history' is more erratic. Compared to other genres, such as 'amusement', 'cycle of life', 'groups', and 'children', it shows a much wider spread in song lengths. It should be noted that in topic modelling the length of a text might influence its topic probabilities: because a long text contains more words, chances are it contains more topics than a short text.

One consequence of the piecemeal, project-based process through which this database has been constructed is a skewed representation of certain epochs. For example, while every Dutch monophonic song dated before 1600 and known to exist is included in the database, coverage of later periods is less complete. We also have to take into account that there is likely a genre bias to be uncovered in the compilation of songbooks – songs did not have similar chances of survival. Another drawback is that even when all known songs from a certain area are entered in the database, this probably is just a snippet of all songs that have been produced during that time. We can assume that a large fraction of the early modern song production is nowadays unknown to us, either because the documents no longer exist, or because they have not been recovered yet. Furthermore, the survival chances of a printed song are considered to be higher than that of a song in manuscript, and that of eighteenth-century song higher than that of a sixteenth-century song. In recent

27 I have used a cut-off length of 1,000 words, since there was only a very small number of songs with a bigger length.

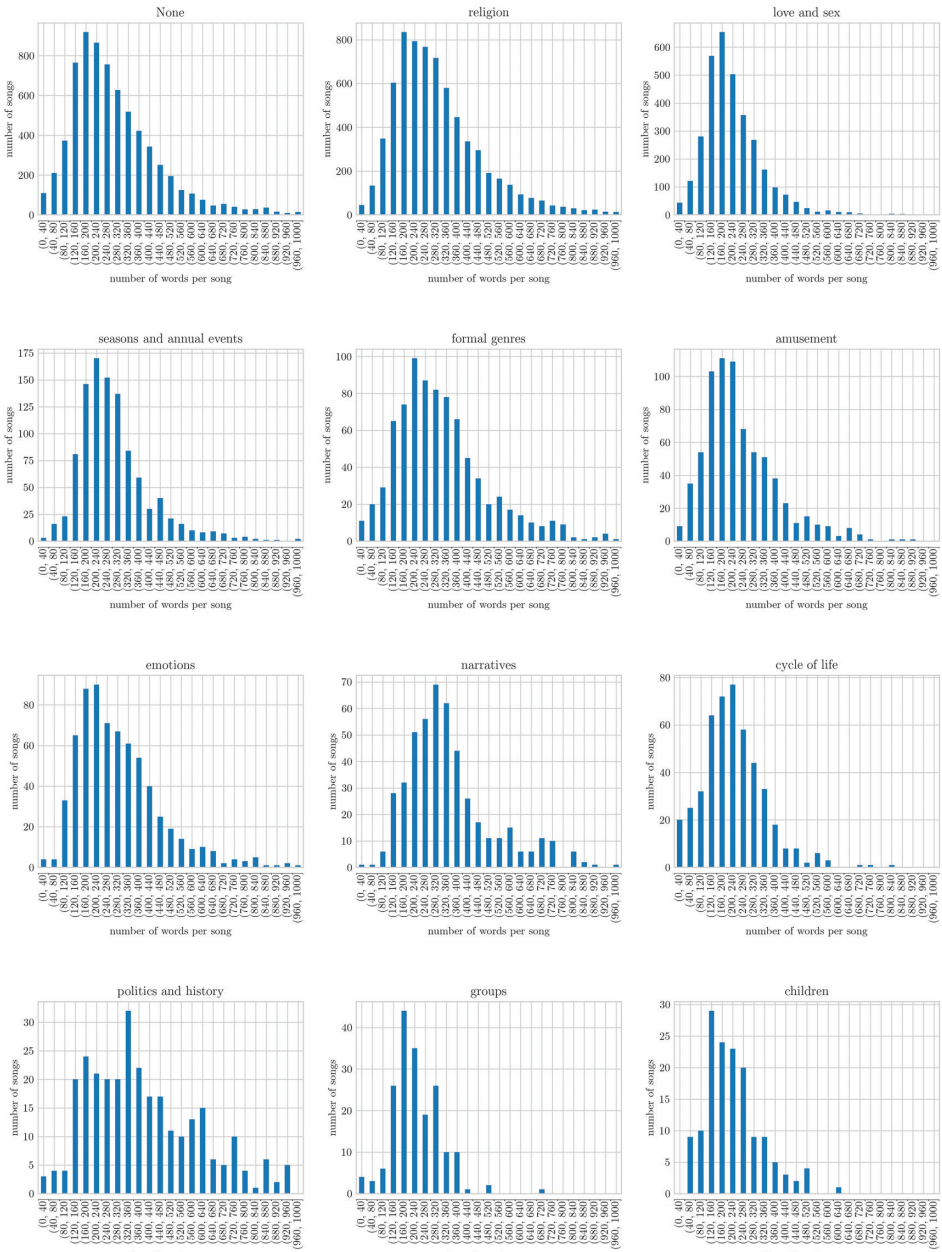

Fig. 2 Song lengths by genre in the Dutch Song Database. N = 22,297.

years, scholars have further explored to what extent the level of loss of historic works can be estimated. Leo Egghe and Goran Proot, for example, have proposed a probabilistic model, based on the frequency with which retrieved copies of historic works survive.[28] Mike Kestemont and Folgert Karsdorp have used an unseen species model from ecodiversity to estimate the survival rate of Middle Dutch chivalric epics.[29] Although this article will not contribute to the study of survival chances of early modern songs, it is important to note that the corpus I have used is undoubtedly biased and incomplete. Furthermore, these biases potentially change over time, and in that case also have implications for a comparative analysis of topics over time.[30]

Topic Modelling

Topic modelling is a way of extrapolating backwards from a collection of documents to infer the discourses that might have generated them. It is assumed that each document in a collection of documents is constructed from a mix of some set of possible topics. A topic can be understood as a collection of words that have different probabilities of appearance in passages discussing the said topic. The model assigns high probabilities to words and sets of words that tend to co-occur in multiple contexts across the corpus. It must be noted that a topic is a probability distribution over the entire vocabulary, and that these distributions can vary hugely between topics. The ten most frequent words of one topic might together make up fifty percent of the topic, while the ten most frequent words of another topic only cover five percent of the topic. Although looking at the ten most frequent words of a topic is common in most topic modelling papers, it is in fact more insightful to look at the number of words of which the sum of probabilities equalizes a certain percentage.

Topic modelling is an unsupervised method, which means that the model makes 'a wager that patterns in the data are sufficiently strong that different latent classes of observations will make themselves "visible"'.[31] This means that the algorithm is not informed about the kind of texts that are imported and what themes it has to look for. The algorithm infers information about individual word meanings based on their repeated appearance in similar contextual situations.

For the purposes of this study, topic modelling was performed using the Python package gensim, which contains a wrapper for MALLET, a Java topic modelling toolkit. I used the statistical technique Latent Dirichlet Allocation (LDA), which is implemented in MALLET as well as in other software packages.[32] LDA is the most commonly used topic modelling method. 'Latent' refers to the concept of latent distributions in the model, and 'Dirichlet' is the name of the most commonly used prior distribution, named after the

28 Egghe and Proot, 'The Estimation of the Number of Lost Multi-Copy Documents'; Proot and Egghe, 'Estimating Editions on the Basis of Survivals'; Proot, 'Survival Factors of Seventeenth-Century Hand-Press Books'.
29 Kestemont and Karsdorp, 'Estimating the Loss of Medieval Literature'.
30 The current study only contains comparative analyses between topics belonging to the same genre.
31 Karsdorp, Kestemont, and Riddell, *Humanities Data Analysis*, 267.
32 Blei, Ng, and Jordan, 'Latent Dirichlet Allocation'.

German mathematician Johann Dirichlet. It consists of a nested multilevel structure, in which the three levels – word, document, and corpus – are distinguished.[33]

Since topic modelling is an unsupervised method for automatically extracting topics from unlabelled data, it can give no guarantees on the reliability and interpretability of the output. A method to measure the coherence of the topics is therefore highly desirable. A coherent topic is a topic that is semantically interpretable: if a topic consists of the words *Jezus*, *kruis* (cross), *lijden* (suffering), and *Golgotha*, we are clearly dealing with the crucifixion of Christ. However, a topic can also be a result of either the model's design settings, or of artefacts in the corpus, such as selection criteria leading to an imbalance in represented genres, resulting in topics in which this imbalance is reproduced. This can (but does not necessarily have to) result in a topic that is not semantically interpretable. An example would be a topic consisting of the words *kribbe* (manger), *prins* (prince), *Holland*, *geus* (beggar), *Venus*, and *minne* (love). A coherence measure indicates whether the formed topics are indeed semantically interpretable.

In a 2015 study, Michael Röder, Andreas Both, and Alexander Hinneburg evaluated the performance of several coherence measures. They concluded that the best performing coherence measure was a rather unexplored one, which they labeled C_v.[34] C_v is based on four parts: i) segmentation of the data into word pairs; ii) calculation of word or word-pair probabilities; iii) calculation of a confirmation measure that quantifies how strongly a word set supports another word set; and iv) aggregation of individual confirmation measures into an overall coherence score.[35] The coherence score is a value between 0 and 1. In general, a higher coherence score indicates a better quality of topic model.[36] To determine the settings for some parameters that have to be set in building a topic model, I used C_v as coherence score.[37]

In order to build a topic model with gensim, a few preliminary steps have to be taken. An important first step is to reduce spelling variation. The early modern period lacked standardised spelling and grammar, which meant that (and was proliferated by the fact that) Dutch authors and printers attached little value to consistency in spelling, resulting in the coexistence of multiple variants between, and sometimes within, texts. Because topic modelling is based on the position of a unique word in a text (and thus each unique spelling variant) and its distribution through a corpus, spelling variation disturbs the results of the process. Reducing the variation in spelling is therefore a crucial step that needs to be taken. For the English language, several tools have been built over the last years to convert historical spelling variants into their modern equivalent. For the Dutch language, however, such options are limited, and the tools that are available come with both advantages and disadvantages. For the purpose of this study, I used the Dutch setup of the VARiant Detector (VARD), a semi-automatic tool introduced by Alistair Baron and

33 Pritchard, Stephens, and Donnelly, 'Inference of population structure'.

34 Röder, Both, and Hinneburg, 'Exploring the Space of Topic Coherence Measures'.

35 For a more comprehensible explanation of the method, see Syed and Spruit, 'Full-Text or Abstract?'.

36 For a detailed study on giving meaning to coherence values of topic models, see Uglanova and Gius, 'The Order of Things'.

37 The default parameters were used in all cases in which no parameters are reported.

Paul Rayson in 2005.[38] The tool uses two lists (a normalized word list and a variant list) to suggest or replace variant words with their normalized counterparts. The normalization suggestions are given based on a combination of four different methods: i) known variant replacements; ii) character edit distance; iii) letter rules; and iv) phonetic distance. The Dutch version of VARD2 was built by Ivan Kisjes and Tessa Wijckmans, and was the result of a collaboration between the CREATE-project of the University of Amsterdam and the Dutch digital research platform *Nederlab*. As a training set, they used the first two books of the 1657 edition of the Dutch translation of the Bible.[39]

VARD2 will find numerous potential normalizations for a given variant. Each of these suggestions is given a confidence score, in the form of a percentage expressing how certain the tool is that this is the correct normalized word. The suggested normalization with the highest confidence score is used to replace each variant. Consider the opening lines of a song from 1598: 'Reyn maecht eerbaer/aenhoort myn claeghen snel' ('Pure honourable virgin, hear quickly my complaint'). VARD2 suggests normalizations for every word, except the word *snel* (quick). The word *reyn* (pure) gets as most popular suggestions the words *rein* (98.47 percent) and *reen* (0.39 percent). The percentages show that in this case, *rein* (clean) is clearly the correct normalization suggestion. For the word *maecht* (virgin), however, the tool offers several potential candidates. The words *maagd* (virgin, 95.59 percent), *macht* (power, 42.71 percent), and *maakt* (makes, 38.52 percent) are suggested as normalizations, although the high percentage corresponding to *maagd* shows that the tool is still confident that this should be the correct normalized word.

However, if the system's normalization methods struggle with a particular variant, the highest confidence score may be relatively low – in these cases a threshold is required, which is the minimum confidence score needed for normalization to take place. If the threshold is not met by the top normalization suggestion, the word will not be normalized, and will be left as a variant. The word *claeghen* (to complain) from the earlier-cited song gets as suggested variants *clean* (0.96 percent) and *clan* (0.93 percent). Clearly, *klagen* is not included in the normalized word list. If the threshold is higher than one percent, the word 'claeghen' will remain unnormalized. A higher threshold will increase precision but reduce recall. After using a single text and checking the number of words that were normalized at different thresholds, I decided to set the threshold at fifty percent. Before normalization, the corpus consisted of 217,906 distinct words; after normalization with VARD2 the number of unique words in the corpus was reduced by 23,857, resulting in a corpus with 194,049 distinct words.[40]

The next pre-processing steps included tokenization – meaning that each text was converted to a list with separate words, or tokens – and punctuation removal. Subsequently, stop words, which are the very common words (in this case mostly function words such as personal pronouns, propositions, and conjunctions), were removed

38 An updated version, the VARD2 tool, was used in this study.

39 Kisjes and Wijckmans, 'Adapting a Spelling Normalization Tool'.

40 In my master's thesis, on which this article is based, I compared the results of three versions of the corpus: the original version, the version that was normalized with a straightforward script that used INL to normalize words, and the version that was normalized with the VARD2-tool. Both the results of the normalization and the coherence of the topic model showed that the latter version of the corpus had the greatest research potential.

from the corpus. Using the Python package nltk, a list of one hundred and fifty most frequent words was made. Content words were manually removed from this list and therefore kept in the corpus.[41] Furthermore, a threshold was set for words that appear too infrequently in the corpus (no_below). Calculating the coherence score of topic models with different settings for no_below, this threshold was set at 2, which means that words that appear in only one document are ignored.[42] The third step was deciding how many topics the model would be creating. Here the coherence score was used again. I built ten topic models with different values for the number of topics, in the range 10:100. The coherence score of each model was calculated. It turned out that the model with fifty topics obtained the highest coherence score, and therefore I decided to build the final topic model using this setting.[43]

The output of a topic model consists of two dataframes (tables): one with the keys of a topic model, the other one with the composition. The dataframe 'keys' contains fifty rows, one for each topic, and two columns: one for the topic id, the other for the words that form a specific topic. The dataframe 'composition' contains as many rows as there are documents in the corpus. The columns are numbered from 0 to 49, corresponding with the id of the topics from 'keys'. The cells from each row contain a value between 0 and 1. The higher a value, the more dominant that topic is in a document. Note that the sum of each row will always be 1, but that the distribution always differs.

After building the final topic model, another tool was used to measure the quality of the model, and to get more insight into the topics and how they are related to each other. I used the Python package pyLDAvis, which is a library for interactive topic model visualization. The package extracts information from a fitted LDA topic model to inform an interactive web-based visualization. After a few steps of pre-processing, the output is an Intertopic Distance Map. Fig. 3 gives an example of the data visualized, in this case topic 27 ('world & money'). Each bubble on the left-hand side of the plot represents a topic.[44] The larger the bubble, the more prevalent the topic is within the corpus. A good topic model should have bubbles scattered throughout the chart, rather than clustered in a single quadrant. A model with too many topics will typically have many overlapping small-sized bubbles clustered in one region of the chart. When hovering over the topics, the thirty most prevalent words of that topic are visualized on the right-hand side. Since a topic model only produces bags of words as output, users have to manually interpret the words that make up a topic. To do so, I looked at the number of words that together made up fifteen percent of a given topic. I built a word cloud from every topic, visualizing the upper fifteen percent of words and their weights within a topic. The top thirty words and their weights within

41 These were the words *God*, *Gods*, *leven*, *hert*, *ziel*, *liefde*, *tijd*, *vreugd*, *here*, *geest*, *woord*, *lief*, *min*, *kwaad*, *zonden*, *heren*, *lof*, *hand*, *wereld*, *mens*, *mensen*, and *recht*.

42 The setting of this parameter was decided after calculating the coherence score of three topic models with a no_below of respectively 2, 5, and 10. The topic model of no_below = 2 gained the highest coherence score.

43 The coherence score of the final topic model was 0.4962.

44 Note that pyLDAvis starts counting at 1, while the topics made with gensim start at 0. Each number *x* in the pyLDAvis-plot therefore corresponds with topic id *x*-1. Moreover, because pyLDAvis starts counting at 1, but gensim at 0, one should add +1 to every topic. Topic 27 is therefore depicted as topic 28 in this visualization.

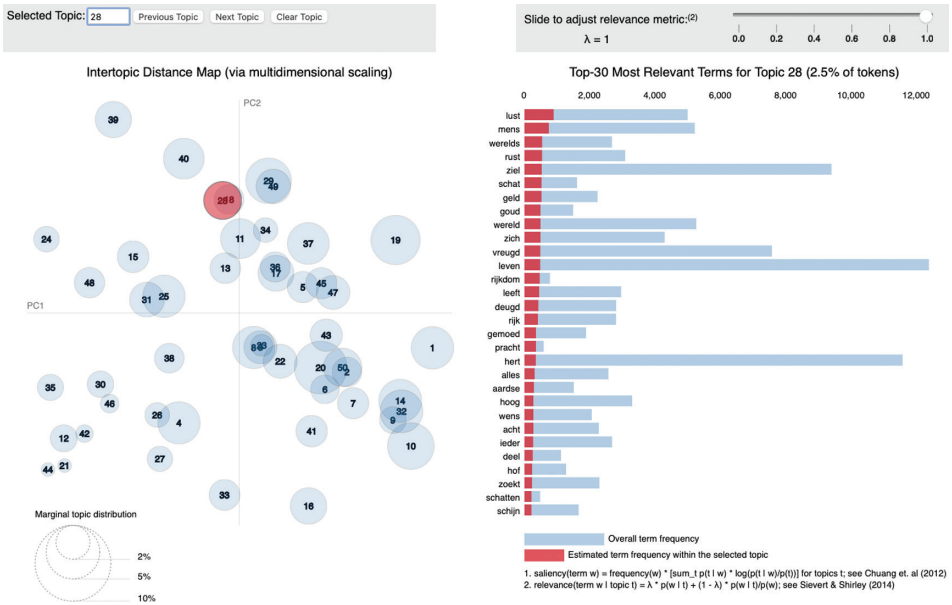

Fig. 3 Visualization of topic 27 ('world & money') in the Dutch Song Database, 1550-1750, using pyLDAvis.

every topic are included in an online appendix.[45] The topics and the subjects I assigned to them can be found in tab. 2.

Dominant Topics

The appendix also includes a column with the sum of the weights of the top thirty words. The lower this value, the 'flatter' the distribution of words within a topic, suggesting that the words in these topics are not typical for that specific topic, but occur in many other topics as well. The clearest example of such a word is 'God': this word has the highest weight in ten topics. Still, its probability can be fairly low (0.00992 in topic 7, 0.0131 in topic 28). An example of a topic with a high sum of weights of the first thirty words is topic 29 (0.2493), to which I assigned the label 'drinking', due to words such as *wijn* (wine), *drinken* (to drink), *vrolijk* (merry), drinkt (drinks), and *bier* (beer). Another clear example is topic 46 (0.2620), to which 'life & sin' was assigned, because of the presence of words such as *wereld* (world), *God*, *leven* (life), *kwaad* (evil), and *sterven* (to die). Topic 5 was labelled as 'Christmas', due to words as *kindeken* (baby), *stal* (stable), *Bethlehem*, and *Maria*. I assigned 'love & sadness' to topic 39, because of the presence of positive words such as *lief* (sweetheart), *liefde* (love), *min* (love), and *schone* (beautiful) on the one hand, and negative words such

45 See for the appendix: https://github.com/awlassche/dutch-historical-song-topics/blob/main/data/appendix.csv (accessed on 18 October 2022).

Tab. 2 Dominant topics in the Dutch Song Database by sum of weights (>= 0.1).

Topic no.	Subject	Sum of weights (> 0.1)
9	religion & old spelling	2847
31	religion	1684
0	religion & life	1508
18	religion	1357
19	verbs	1312
39	love & sadness	1172
13	religion	1096
38	love & happiness	1072
28	prayer & preaching	1026
24	rejection	997
10	suffering & sadness	799
2	love & tragedy	788
6	possessives	782
36	religion	776
1	religion & Mary	776
3	verbs	769
30	love	761
48	love	739
27	world & money	738
5	christmas	735
7	God & enemy	715
49	God & country (praise)	691
32	nation & country	670
46	life & sin	642
12	marriage	635
8	life & praise	631
44	religion & virtue	624
15	Old Testament	614
47	bucolic	613
21	love & happiness	591
42	Good Friday & Easter	571
16	good & evil	567
14	myth & beauty	559
23	physical love	533
40	New Testament	529
17	heaven & happiness	524
4	religion & happiness	509
34	seduction	479
29	drinking	465
37	nature	433
11	unclear	345
26	money & work	307
33	unclear	285
35	religion	264
45	old or odd spelling	259
25	sea	248
22	church	203
41	German	171
43	French	119
20	solfege	92

Source: Dutch Song Database.

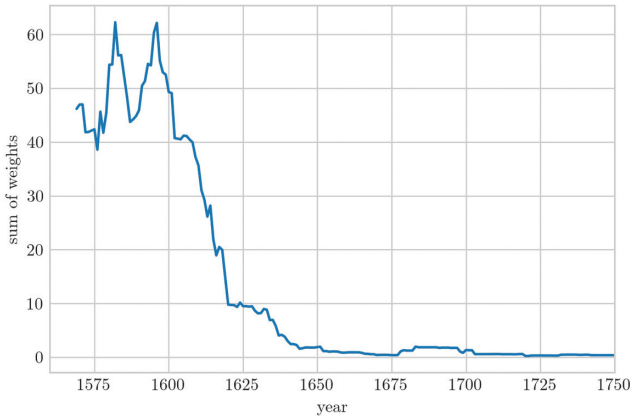

Fig. 4 Relative frequency of topic 9 in the Dutch Song Database, by annual sum of weights (>= 0.1).

as *ach* (oh!), *pijn* (hurt), *smart* (grief), and *verdriet* (sorrow) on the other. This topic differs from another topic on love (topic 38) in that almost all the words in this topic are positive: *hart* (heart), *vreugd* (joy), *leven* (life), *min* (love), and *vermaak* (entertainment). I assigned 'love & happiness' to this topic.

I labelled topic 32 'nation & country': the words *prins* (prince), *stad* (city), *land* (country), *graaf* (count), and *Holland* leave no doubt that these are words from political songs. However, the sum of weights of the first thirty words in this topic is rather low: 0.1342. The same goes for topic 14. While the words all refer to a topic that I labelled 'myth & beauty' (including *Venus*, *schoonheid* (beauty), *gezicht* (face), *glans* (shine), and *goden* (gods)), the sum of weights is only 0.1429. This suggests that these topics have a flat distribution. In topic 47, words from pastoral or bucolic songs are clustered. Many words refer to the shepherd's life, such as *vee* (cattle), *herder* (shepherd), *bos* (forest), and *schaepjes* (little sheep). I assigned the label 'bucolic' to this topic. Topic 28 was labelled 'prayer & preaching', because apart from the obvious religious words, there are words in this topic that have an oral connotation – *hoort* (hears), *horen* (to hear), *zeid* (said), *verstaat* (understands), *vermaan* (admonish), and *broeders* (brothers). It turns out that many topics have something to do with religion. This is not surprising, since we already know that almost forty percent of the corpus are songs from the category 'religion'. In some cases, it is quite easy to define a label for a topic (such as 'religion & happiness', 'religion & sin'), but in other cases the differences are more difficult to distinguish.

To examine which topics dominate the corpus, I selected for each song the topics that covered at least ten percent of the song, meaning that a topic has a minimal weight of 0.1 in a song. Subsequently, I summed these weights per topic (see tab. 2). The topics are ranked by the sum of their weights. Topic 9 is by far the most popular within the corpus. The appendix shows this topic has a rather 'flat' probability distribution: the five most frequent words in this topic (*god*, *wt*, *leun*, *gods*, *nv*) all have a weight of between 0.01 and 0.02. The oddly spelled words such as *wt*, *leun*, *nv*, *si*, *bouen*, and *ouer* suggest that topic 9 is rather a result of artefacts in the corpus: the tool used for spelling normalization does not work well on texts written before 1637. Fig. 4 confirms this hypothesis: topic 9 peaks between 1550 and 1600, but descends soon after.

The rest of the top ten of most dominant topics is made up by topics that relate either to religion or to love, except topic 19, in which words are not clustered on semantic similarity, but on linguistic characteristics. I labelled this topic as 'verbs', because it consists of words such as *leven* (to live), *zullen* (shall), *geven* (to give), and *komen* (to come). The relatively low sum of weights of the first thirty words (0.1573) indicates that these words are not very distinctive for this particular topic, but occur in many other topics. If we take a look at the bottom of the chart, the opposite is the case. Topic 41 consists of German words (*das* (the), *mein* (mine), *mir* (to me), *ein* (a/an), *mit* (with)) and has a sum of weights that is above average (0.2697). The same goes for topic 43, which consists of French words and has a sum of weights of 0.3463, as well as for topic 20, consisting of interjections that often occur in children's or folklore songs, with a sum of weights of 0.5108. These high values indicate that these words are rather distinctive for these topics.

The most dominant topics in the corpus, then, are 'religion plus something'. This is not a surprising result, because the distribution of categories already indicated many religious topics. Still, the topic model has uncovered many differences within the wide-ranging genres of songs. Although many more observations can be made with regard to the fifty most distinctive topics, the next two sections will explore how the dominance of song topics fluctuates in the context of the two cases I introduced earlier: Petrarchan songs and patriotic songs. As we have seen, these are topics that are popular within the corpus, and based on qualitative research we would expect to see some changes over time regarding their popularity. A quantitative approach can be used to detect this.

Petrarchism Songs

Based on the existing literature, song topics related to the Petrarchistic tradition should be peaking in the beginning of the seventeenth century. These could be topics on the tragic aspects of love, but also topics which can be related to the Renaissance and classical tradition, for example the bucolic genre. Two dominant topics in the second-largest overarching category – love – are contradictory: topic 39, 'love & sadness', and topic 38, 'love & happiness'. Fig. 5 shows the evolution of these two topics over time, which reveals an interesting pattern: for instance, the topic 'love & sadness' peaks about fifty years before the topic 'love & happiness'.[46] Initially, topic 39 is dominant, but by 1700, topic 38 ('love & happiness') picks up and rapidly becomes the dominant of the two love topics. The topic 'love & sadness' does increase slightly in the eighteenth century, but never comes close to 'love & happiness'. 'Love & sadness' can be seen as a typical Petrarchan topic: the words *ach* (woe!) and *pijn* (pain) in this topic speak for themselves. Fig. 5 demonstrates that the popularity of these songs peaked in the early seventeenth century, as expected. It also suggests that towards the end of the seventeenth century, songs with negative emotions were replaced by songs with a rosier view on love.

46 I used a rolling average of twenty years in all graphs, to let 'accidental' differences between years have less impact on the analysis.

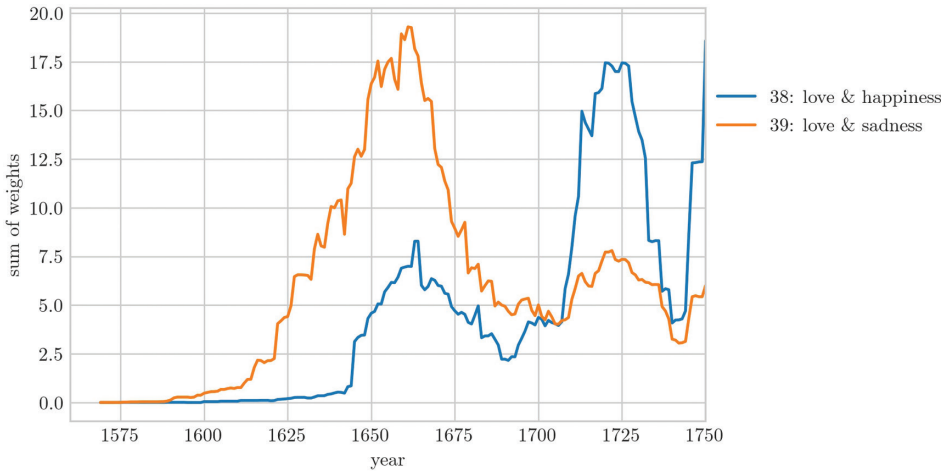

Fig. 5 Relative frequency of topics 38 and 39 in the Dutch Song Database, by annual sum of weights (>= 0.1).

Fig. 6 shows additional love topics besides the two discussed above, demonstrating that another Petrarchan topic (2, 'love & tragedy') had in fact peaked in the first quarter of the seventeenth century, before being replaced by topic 39. When we take a closer look at the words that form the latter two topics, it becomes clear that they exhibit many similarities: both contain the words *liefste* (dearest), *liefde* (love), *herte* (heart), *minnen* (to love), *pijn* (pain), and *verdriet* (sorrow). However, both topics also contain distinctly different words: topic 2 includes words such as *troost* (consolation), *lijden* (suffering), *genezen* (to heal), and *scheiden* (to separate), while topic 39 consists of words such as *smert* (grief),

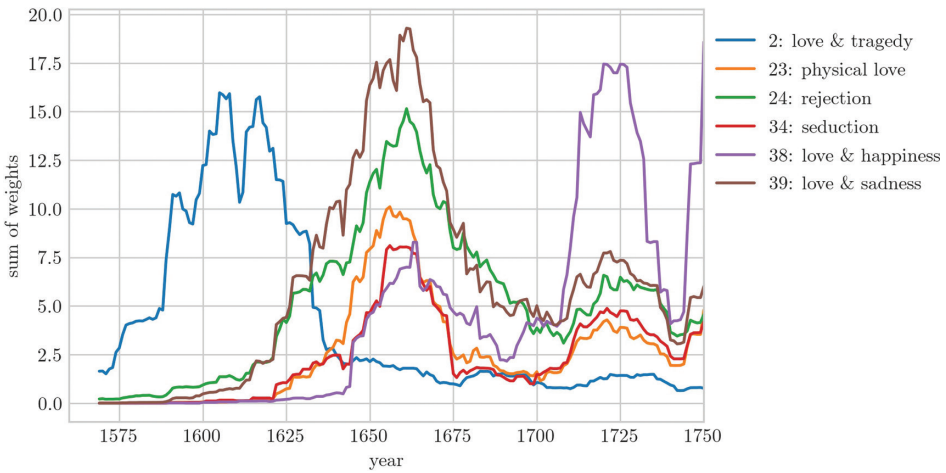

Fig. 6 Relative frequency of topics 2, 23, 24, 34, 38, and 39 in the Dutch Song Database, by annual sum of weights (>= 0.1).

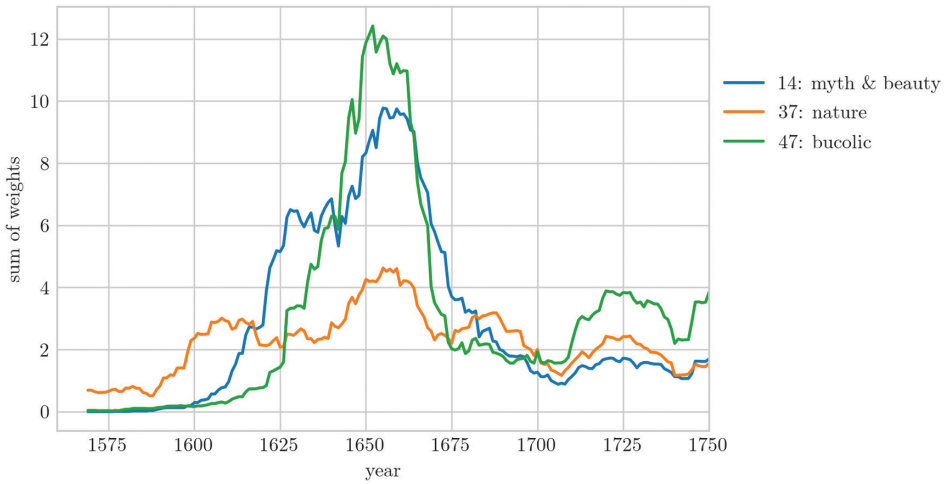

Fig. 7 Relative frequency of topics 14, 37, and 47 in the Dutch Song Database, by annual sum of weights (>= 0.1).

waarom (why), *droeve* (sad), and *sterven* (to die). Semantically the topics are related, but they include different words. Taken together, the two topics suggest the dominance of tragic love over all other love tropes in songs published between 1575 and 1685.

Topic 24 ('rejection') peaked simultaneously with 'love & sadness', between 1650 and 1675. Other love variants, such as 'seduction' and 'physical love', are less prevalent, but still peak at the same time as topics 39 and 24. These topics may also be connected to the heyday of Petrarchism. A famous Dutch imitator of Petrarch was Pieter Corneliszoon Hooft, whose poems were largely based on the Petrarchan exemplar, in particular the conventions of amorous lyric. The case of Hooft demonstrates that not only poems but also songs were written with the Petrarchan tradition in mind. Hooft's songs ended up in songbooks, accompanied by melodies, and reached a large audience. From 1607, song poets often referred when assigning their melodies to songs from Hooft's play *Granida* (1605).[47] These observations are reflected in fig. 6, which shows that various love topics, which in their own way all relate to Petrarchan love poetry, dominated between 1575 and 1650.

Other topics related to the humanistic tradition are 'bucolic', 'myth & beauty', and 'nature', which are visualized in fig. 7. 'Bucolic' refers to the pastoral, a genre originating in the classical tradition and especially in the works of Theocritus and Virgil. The use of mythological figures such as Venus and Phoebus in topic 14 also indicates the imitation of classical genres. There is a clear decline in these topics after 1660. 'Love & sadness', 'love & tragedy', and 'rejection' are overtaken by 'love & happiness' (fig. 6). The Petrarchan tradition with negative emotions on love seems to have moved to the background, making room for a less dramatic, more light-hearted interpretation of love. An explanation could be that the moralistic attitude described by Keersmaekers, which can be linked to the

47 Hooft, *Granida*, 1.

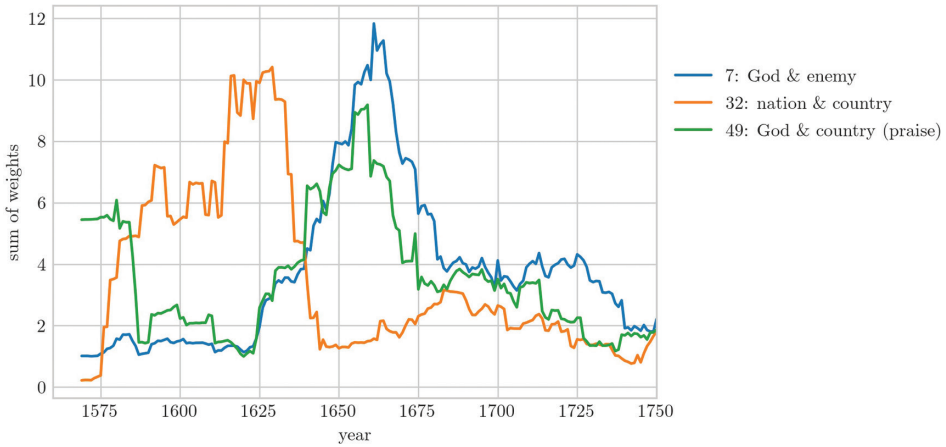

Fig. 8 Relative frequency of topics 7, 32, and 49 in the Dutch Song Database, by annual sum of weights (>= 0.1).

culture of piety that was blooming in the seventeenth century, was beginning to diminish. Qualitative research could provide more insight into whether (and if so, why) the depiction of the Petrarchan tradition in songs declined in the course of the eighteenth century.

Patriotic Identity

Scholarship has suggested that topics dealing with nation and country peaked during the Dutch Revolt, although they remained dominant after the 1648 Peace of Münster. I have selected three topics related to politics to test this hypothesis. Fig. 8 shows that topic 32 ('nation and country') appeared around 1575 and steadily rose thereafter. The topic remained dominant for at least fifty-five years, a timespan largely coinciding with the Dutch Revolt. Dominant words in this topic are *prins* (prince), *land* (country), *graaf* (count), *spaensche* (Spanish), and *paus* (pope), confirming that these topics are indeed covered by the beggar songs, which were extremely popular during the Dutch Revolt. It seems likely that the many reprints of the *Nieu Geusen liedtboecxken* (published between 1577 and 1687) are responsible for the dominance of this topic. The song that is most emblematic for topic 32 (because of its high weight) is titled 'Wel op wel op Spangiaerden', an anti-Spanish song from the *Nieu Geusen liedtboecxken*.[48] However, more songbooks with patriotic songs existed in the Dutch Republic.[49] Although some of the songs in these songbooks were also included in the beggar songbooks, others that would have fitted the criteria were not. Topic 32 shows a drastic fall around 1640, which coincides with the Peace of Münster in 1648. The topic is no longer dominant in subsequent decades, bar a little revival between 1675 and 1725. Since all graphs use a rolling average over twenty

48 *Een Nieu Geuse Lieden boecxken*, no. 116.
49 Grootes, 'Liedjes over de Tachtigjarige Oorlog', 173.

years, this revival could be the result of the last reprint of the *Nieu Geusen liedtboecxken* in 1687. All in all, fig. 8 suggests that the dominance shown by this topic during the Dutch Revolt did not return, merely that there is perhaps a renewed popularity of patriotic songs during the War of the Spanish Succession (1701-1714), though the line's erratic course demands that caution be shown in making such an interpretation.

Around 1625, when topic 32 began to decrease, the two others gained in popularity, of which 'God and enemy' became even more dominant than topic 32. An important contrast between topics 7 and 49 on the one hand, and topic 32 on the other, is the presence of religious words in topics 7 and 49. The only religious word in topic 32 is *God*, with a considerably lower weight, given that words such as *prins* (prince), *land* (country), *stad* (city), and *vijand* (enemy) are more prevalent. Topics 7 and 49 contain more overlap with other religious topics, with the presence of such words as *God*, *boze* (angry), *Gods* (God's), and *handen* (hands) in topic 7, and *God*, *zijne* (His), *naam* (name), *heren* (the Lord's), and *hemel* (heaven) in topic 49. This suggests these are religious songs with a strong focus on the aspects of war, peace, enemy, and the people. The song with the highest weight contribution to topic 49 turns out to be a psalm in a translation by Petrus Datheen; indeed, many other songs with a high weight contribution to topic 49 are also psalms. This is not surprising, as many psalms use a vocabulary that expresses a battle against a real or imaginary enemy. The song with the highest weight contribution to topic 7 is from the songbook *Stichtelijcke Gesangen* (1661) and relates the story of the biblical figure Daniel, who was thrown into a lion's den:

> When Daniel came in the den
> Of cruel lions
> (Who howled with a loud
> Cry of ferocity),
> As if into the abyss,
> Surrounded by wailing,
> The angel, by God's command,
> Held the lions' jaws,
> So that they sat meek
> As if they were lambs.
>
> The entire world is crawling
> With cruel enemies,
> Who would have defiled us,
> If God had not,
> By his great might,
> Chained their hands:
> But now he puts a limit
> To all the tyrant's work:
> And takes heed of his children,
> Strong and all-knowing.[50]

Although the first verse relates Daniel's well-known fate, the second contains a reference to a war the singer was involved in. This not only ties in with the aforementioned increase

50 Maertsz., *Stichtelijcke gesangen*, 200-201.

in attention paid by orthodox ministers to patriotic ideas, but also with a broader trend in orthodox circles: during the Further Reformation (a Pietistic movement that began in the second half of the seventeenth century), songs were increasingly considered to be a literary genre through which the dissemination of Pietistic ideas could take place. Proponents of the Further Reformation published their own catechisms, sermon collections, and songbooks, which caused an exponential rise in the number of Reformed songbooks after 1650.[51]

Another explanation for these trends could be that these topics, in addition to their political connotation, also have a military meaning. Furthermore, the topics might also be (partly) formed of songs in which earlier military events are historicized. In summary, the topics related to the fatherland not only show how these songs prevailed during political conflict, but the patriotic ideas they embodied also became increasingly embedded in the seventeenth-century Pietistic tradition.

Conclusion

This article has shown that computational methods for text analysis offer a meaningful way to analyse a large digital corpus. The benefits of this study are, therefore, multiple: first and foremost, topic modelling has provided a possible solution to the issues addressed in the introduction – the lack of metadata on genres in the Dutch Song Database, and the impossibility of searching for genre patterns in this collection. Although the dominance of religious and love topics was to be expected, given the existing genre classification of the database, this study shows that subcategories can be defined within a genre. For example, the built topic model was able to distinguish between several aspects of love, different focuses within the topic 'religion', and differences in songs with political connotations. These results contribute to a better understanding of the richness of the Dutch Song Database. A recommended next step would be to supplement the missing metadata on genres in the database with the results from this study, as this would facilitate its use by other researchers.

Secondly, by exploring the relationship between topical fluctuations and cultural-historical changes in the Dutch Republic, I have shown how quantitatively obtained results can be related to existing scholarship on early modern Dutch song culture. This study confirms that Petrarchan ideas were indeed transmitted through song, a phenomenon that began in the early seventeenth century. Furthermore, the results suggest that the Petrarchan tradition in songs started to decline in the course of the eighteenth century. In addition, this study shows that songs with a patriotic identity became more popular during political crises, and that songs belonging to the religious genre were also used to spread patriotic ideas, a trend coinciding with the heyday of the Further Reformation.

More generally, I have shown how, with the methodology used in this study, diachronic developments in textual corpora can be explored. This particular study shows how song lyrics interfere with and react to socio-cultural and political developments,

51 Porteman et al., *Een nieuw vaderland voor de muzen*, 658-660.

thus demonstrating that songs are a suitable medium through which the historian may gauge the social climate. The question rises, however, whether songs are precursors to or merely the result of socio-cultural trends, something that deserves further research. Furthermore, the proposed method should be improved, so that it may be applied to other corpora, both early modern and modern, both in Dutch and other languages.

A few remarks should be made regarding the methods used and the results obtained. In the first place, this study shows the complexity of using historical textual data on a large scale. Although retaining spelling variation might be necessary for other studies or disciplines, in this case it was important that spelling be normalized across the corpus. The method used to reduce spelling variation in this study is not flawless. For example, because the training set of the tool was built from seventeenth-century material, contemporary words were not normalized by the tool. This resulted in topics that were not semantically interpretable; rather, a group of word clusters formed that were based purely on the fact that their words were not normalized. Clear examples are topics 9 and 45, which contain words such as *dij*, *yn*, *soe*, *uwz*, *jon*, and *fen*. This shows once again that more research into techniques to deal with historical variation in Dutch spelling is crucial.

It must also be noted that topic modelling is an oft-debated computational method, not only because it is based on probabilities and the mathematical idea behind the method seems a mystery to many users, but also because it is clear that more work is required on measures to evaluate the results. In this study, evaluation was done by using coherence measures and visualizations, but it is important to stress that these evaluation methods are also not flawless. Another point where intervention of the researcher may have led to biased results was the assigning of a description to topics. This was done by looking at the word cloud of a topic's fifteen percent most dominant words, but it is entirely possible that another researcher would have assigned different subjects to the topics.

In general, it should be noted that when using topic modelling, songs originating from a certain context, and sung by multiple groups of peoples for various reasons, are simply reduced to groups of words. It is always necessary, therefore, to interpret and evaluate the results of a topic model in detail, as well as put them in the cultural, sociological, and historical context in which the texts were produced. I have tried to do this to a certain extent, but further research can undoubtedly improve my findings. The same goes for other aspects of songs and song cultures that were not taken into account in this study, but that can certainly affect a song's popularity, such as its musical style and communal singing practices.

Scholars have published many qualitative studies on well-known and lesser-known songs, songbooks, and song writers from the Dutch Republic. The aim of this study, in contrast, has been to offer a quantitative perspective, and to demonstrate how the combination of both approaches can offer new perspectives to the field of early modern literary studies. My quantitative research has generated new hypotheses about the Dutch song culture that merit investigation, both by quantitative *and* qualitative literary researchers and cultural historians. I believe that zooming out, while inevitably rendering some details invisible, allows us to see new patterns and generate new hypotheses that can improve our understanding of not only early modern song culture, but Dutch culture at large.

Bibliography

Blei, David M., Andrew Y. Ng, Michael I. Jordan, and John Lafferty (eds.), 'Latent Dirichlet Allocation', *Journal of Machine Learning Research* 3 (2003/4-5) 993-1022.

Bork, Gerrit Jan van, et al., *Algemeen Letterkundig Lexicon* (Leiden 2012).

Bruin, Martine de, 'Het Wilhelmus tijdens de Republiek', in Louis Peter Grijp (ed.), *Nationale hymnen. Het Wilhelmus en zijn buren* (Amsterdam 1998) 16-42.

Een Nieu Geuse Lieden boecxken (Enkhuizen: Jan Jacobs Palensteyn, 1625).

Egghe, Leo, and Goran Proot, 'The Estimation of the Number of Lost Multi-Copy Documents. A New Type of Informetrics Theory', *Journal of Informetrics* 1 (2007/4) 257-268.

Gelderblom, Arie-Jan, 'Investing in Your Relationship', in Els Stronks and Peter Boot (eds.), *Learned Love. Proceedings of the Emblem Project Utrecht Conference on Dutch Love Emblems and the Internet* (The Hague 2007) 131-142.

Grijp, Louis Peter, *Het Nederlandse lied in de Gouden Eeuw. Het mechanisme van de contrafactuur* (Amsterdam 1991).

Grootes, Eddy, 'Het jeugdig publiek van de "nieuwe liedboeken" in het eerste kwart van de zeventiende eeuw', in Willem van den Berg and Hanna Stouten (eds.), *Het woord aan de lezer. Zeven literatuurhistorische verkenningen* (Groningen 1987) 72-88.

Grootes, Eddy, 'Liedjes over de Tachtigjarige Oorlog in andere bundels dan het Geuzenliedboek', *De Zeventiende Eeuw* 13 (1997/1) 173-179.

Hooft, Pieter Cornelisz, *Granida*. Lia van Gemert (ed.) (Amsterdam 1998).

Jensen, Lotte (ed.), *The Roots of Nationalism. National Identity Formation in Early Modern Europe, 1600-1815* (Amsterdam 2016).

Karsdorp, Folgert, Mike Kestemont, and Allen Riddell, *Humanities Data Analysis. Case Studies with Python* (Princeton 2021).

Keersmaekers, August Albert, *Wandelend in Den Nieuwen Lust-Hof. Studie over een Amsterdams liedboek 1602-(1604)-1607-(1610)* (Nijmegen 1985).

Kestemont, Mike, and Folgert Karsdorp, 'Estimating the Loss of Medieval Literature with an Unseen Species Model from Ecodiversity', *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)* (Amsterdam 2020) 44-55.

Kestemont, Mike, et al., *Van wie is het Wilhelmus? De auteur van het Nederlandse volkslied met de computer onderzocht* (Amsterdam 2017).

Kisjes, Ivan, and Tessa Wijckmans, 'Adapting a Spelling Normalization Tool Designed for English to 17th Century Dutch', *Proceedings of the Digital Humanities Conference 2018* (Mexico City 2018) 412-413.

Kranenburg, Peter van, Martine de Bruin, and Anja Volk, 'Documenting a Song Culture. The Dutch Song Database as a Resource for Musicological Research', *International Journal on Digital Libraries* 20 (2019/1) 13-23.

Maertsz., Cornelis, *Stichtelijcke Gesangen, behelsende Bybelsche In-vallen, Geestelijcke Bedenckingen, Eerlycke Vermaeckingen* (Hoorn: Gerbrant and Jan Martensz., 1661).

Poel, Dieuwke van der, Louis Peter Grijp, and Wim van Anrooij (eds.), *Identity, Intertextuality, and Performance in Early Modern Song Culture* (Leiden 2016).

Porteman, K., et al., *Een nieuw vaderland voor de muzen. Geschiedenis van de Nederlandse literatuur 1560-1700* (Amsterdam 2009).

Pritchard, Jonathan K., Matthew Stephens, and Peter Donnelly, 'Inference of population structure using multilocus genotype data', *Genetics* 155 (2000/2) 945-959.

Proot, Goran, 'Survival Factors of Seventeenth-Century Hand-Press Books Published in the Southern Netherlands. The Importance of Sheet Counts, Sammelbände and the Role of Institutional Collections', in Andrew Pettegree and Flavia Bruni (eds.), *Lost Books. Reconstructing the Print World of Pre-Industrial Europe* (Leiden 2016) 160-201.

Proot, Goran, and Leo Egghe, 'Estimating Editions on the Basis of Survivals. Printed Programmes of Jesuit Plays in the "Provincia Flandro-Belgica" before 1773, with a Note on the "Book Historical Law"', *The Papers of the Bibliographical Society of America* 102 (2008/2) 149-174.

Röder, Michael, Andreas Both, and Alexander Hinneburg, 'Exploring the Space of Topic Coherence Measures', *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining – WSDM' 15* (Shanghai 2015) 399-408.

Stronks, Els, 'Identiteiten van adolescenten in de vroegmoderne liedcultuur. Het studentenlied als casus', *Nederlandse Letterkunde* 17 (2012/3) 225-248.

Stronks, Els, '"Dees kennisse zuldy te kope vinnen". Liedcultuur en de waarde van "know how" in de vroegmoderne Republiek', *De Zeventiende Eeuw* 30 (2014/2) 147-167.

Syed, Shaheen, and Marco Spruit, 'Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation', *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (Tokyo 2017) 165-174.

Uglanova, Inna, and Evelyn Gius, 'The Order of Things. A Study on Topic Modelling of Literary Texts', *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)* (Amsterdam 2020) 57-76.

Van der Haven, Cornelis, 'Patriotism and Bellicism in German and Dutch Epics of the Enlightenment', *Arcadia* 47 (2012/1) 54-77.

Veldhorst, Natascha, 'Pharmacy for the Body and Soul. Dutch Songbooks in the Seventeenth Century', *Early Music History* 27 (2008) 217-285.

Veldhorst, Natascha, *Zingend door het leven. Het Nederlandse liedboek in de Gouden Eeuw* (Amsterdam 2009).