ETHNOMUSICOLOGY FORUM
2025, VOL. 34, NO. 3, 411–442
https://doi.org/10.1080/17411912.2025.2588780

Routledge
Taylor & Francis Group

Introducing the digitised dataset of Slovenian folk ballads

Vanessa Nina Borsan [a,b], Mojca Kovačič [c], Mathieu Giraud [a], Marjeta Pisk [c], Matevž Pesek [d] and Matija Marolt [d]

[a]CNRS, Centrale Lille, UMR 9189 CRIStAL, University of Lille, Lille, France; [b]Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia; [c]ZRC SAZU Institute of Ethnomusicology, Ljubljana, Slovenia; [d]Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia

ABSTRACT

This article introduces a Slovenian folk song ballad dataset consisting of annotated transcriptions of 402 monophonic songs, some of which also come with recordings. It traces the historical trajectory of the materials from their collection to their digitisation, providing a statistical overview of the collection's structure and outlining the theoretical framework for the analysis of key elements: context (metadata), content (music), and lyrics. The study acknowledges the significance of historical folkloristic and ethnomusicological practices, integrating these into the synthesis of the materials. It addresses issues such as collector bias, divergent transcription styles, and the inaccessibility of certain recordings.

KEYWORDS

Ethnomusicological archives; digital humanities; Slovenian folk ballads; music annotation; computational analysis; transcription practices

Introduction

Fieldwork, transcription, and data conservation can take significantly different forms in different time periods. Hence, when examining materials collected in the field, the more information we gather about them, the closer we come to knowledge. Floridi (2010) proposes that the latter, knowledge, can be understood as a gradual reduction of what he terms a *data deficit*. In his framework, data in themselves are merely raw, uninterpreted elements. Once these data are organised and given meaning, they become *semantic content*. When such semantic content is also accurate or truthful, it qualifies as *semantic information*. Knowledge, in turn, emerges from the systematic accumulation and interpretation of this truthful information. What constitutes truth remains up for debate. However, the more diverse the information, the more insightful the investigation. The choice of information we collect is a matter of both practical and theoretical concern. Within ethnomusicology, as Timothy Rice points out, various truths can emerge 'from different social and historical positions, interpretations of meaning, plumbing reflexively the depths of individual experience' (Rice 2017: 166). This understanding, if not fully appreciated, can become a source of confusion in ethnomusicological thought (Rice 2017), as it is important to remember that collections are 'a product of their time'

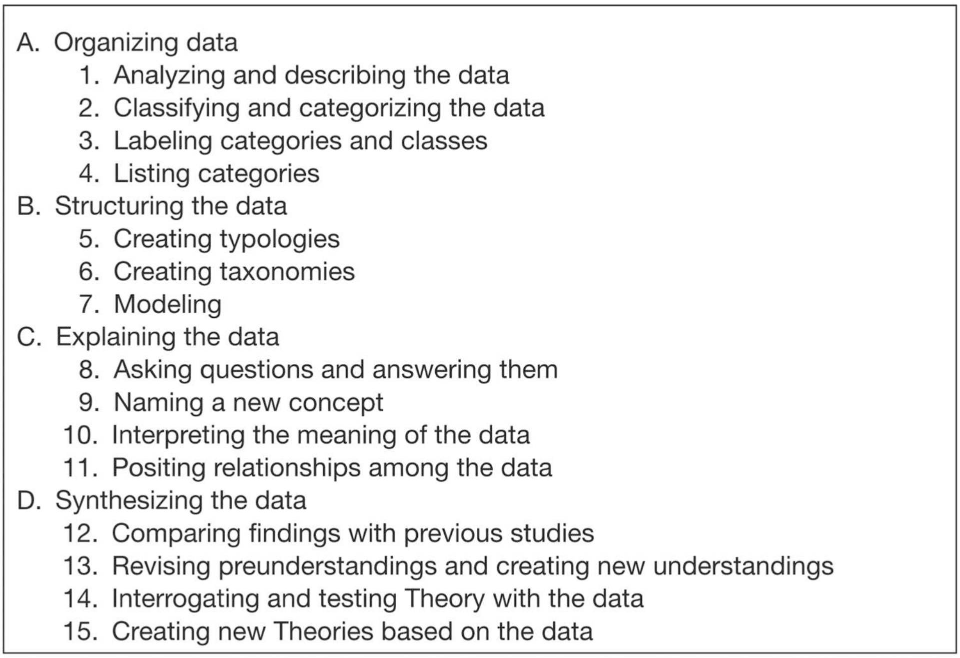
CONTACT Vanessa Nina Borsan vanessanina.borsan@ff.uni-lj.si CNRS, Centrale Lille, UMR 9189 CRIStAL, University of Lille, Lille F-59000, France

(Golež Kaučič et al. 2007) and are therefore collected, transcribed, and analysed within specific historical settings.[1]

Considering this, Rice appears to describe ethnomusicological *theorising* in terms of one initial step followed by four further stages (Rice 2017: 27–9). The first step, *data collection*, is occasionally considered the preface—a precondition that must be fulfilled before research. However, because of the inevitable attribution of meaning to data during collection, Rice understands it as the initial step (Rice 2017). The next four steps—*organising*, *structuring*, *explaining*, and *synthesising* the data—are further divided into 15 activities (Figure 1).

Following Rice's model, this article presents an annotated dataset of Slovenian folk song ballads, including 402 digitised transcriptions (collected between 1819 and 1995) and 22 recordings available on the Dezrann platform and released as a downloadable dataset. The collection under discussion represents a corpus of Slovenian folk song ballads centred on family-related themes, previously curated by members of Ethnomusicology Institute ZRC SAZU. These materials originate from their existing physical collection, already classified under this genre, and are now being digitised, digitally published, annotated, and further analysed.

We begin by situating the dataset within its historical context and then proceed to discuss the theoretical framework and analytical contributions of our work. In the next section, 'Collection and Conservation Practices', we focus on the methods, from fieldwork to digitisation, used to collect and preserve Slovenian folk songs, supported by theoretical frameworks for data collection drawn from Merriam and Merriam (1964), Rice (2017), Lomax (1978), and select Slovenian ethnomusicologists, such as Drago Kunej, Urša Šivic, Marjetka Golež Kaučič and others.



A. Organizing data
   1. Analyzing and describing the data
   2. Classifying and categorizing the data
   3. Labeling categories and classes
   4. Listing categories
B. Structuring the data
   5. Creating typologies
   6. Creating taxonomies
   7. Modeling
C. Explaining the data
   8. Asking questions and answering them
   9. Naming a new concept
   10. Interpreting the meaning of the data
   11. Positing relationships among the data
D. Synthesizing the data
   12. Comparing findings with previous studies
   13. Revising preunderstandings and creating new understandings
   14. Interrogating and testing Theory with the data
   15. Creating new Theories based on the data

Figure 1. Rice's types of ethnomusicological *theorising* (Rice 2017: 30).

Next, we explore methodologies for ‘Organising’ and ‘Structuring’ various folk song materials, with a focus on the processes of digitisation and data annotation. Following this, we turn to issues in the organising and structuring of this particular dataset, broadly categorised into work with metadata, descriptors, and melodic sequences. This dataset comprises 402 digitally visualised Slovenian folk song ballads with annotations, as well as 22 recordings available on the Dezrann platform. Additionally, we provide information on data accessibility. In the final section, we offer ‘Explanation’ of the content, supported by relevant statistical data, and examine the correlations between metadata, descriptors (including on metre, tone sets, and lyrics), and melodic sequences. Then, our ‘Synthesis’ explores the rationale behind the varying degrees of correlation and highlights the potential for misleading interpretations to arise from quantitative analyses, particularly when certain correlations are concerned (for instance, the case of ‘few-tone’ melodies).

Collection and conservation practices (in Slovenia)

In our collection, transcriptions and (later) recordings span from the early nineteenth to the late twentieth century, a period during which notable changes in methods and technologies for material collection influenced their present condition.

The pre-recordings era

Chronologically, the collection starts with songs from a period prior to the advent of sound recording devices. These songs appear primarily in the form of lyric manuscripts taken from various smaller collections (those of Emil Korytko, Oroslav Caf, Anton Breznik, Matija Majar Ziljski, and others). Many of these manuscripts were later incorporated into a monumental collection by Karel Štrekelj (1859–1912), and, subsequently, one by Joža Glonar (1885–1946), who between 1895 and 1923 edited numerous volumes of Slovenian folk songs titled *Slovenske narodne pesmi*. Štrekelj was the first to organise folk songs by theme, assigning a number to each song, and the first to provide rules for lyric transcription.

Štrekelj and Glonar adopted a simplified dialectal transcription, tailored to the abilities of the transcribers and of the general public. This basic dialectal system remains in use to this day. Early collectors, lacking transcription skills and influenced by the collection practices and socio-political goals of the time, did not prioritise melodic transcriptions, rarely including them. Drawing on a relatively detailed definition of a folk song, Glonar completed the collection process by publishing a total of 8,686 songs and 1,944 units of ‘supplementary material’, adding introductory verses or stanzas, as well as items that were previously not considered authentic. Each song title also noted the recording location in parentheses, while other information was included in footnotes (e.g. the recordist’s name, location, date and time, source, and textual variations from unpublished versions). Despite several difficulties and uncertainties—such as deciding what qualifies as a folk song and dealing with missing or inconsistent metadata—he nevertheless contributed to the establishment of systematic collection standards.

The earliest transcriptions made by collectors prior to Glonar and Štrekelj are regarded as unreliable by contemporary standards. Franjo Kuhač (1834–1911) for instance, collected folk melodies but presented them in arranged editions for voice and piano

(Kuhač 1878–1881), thereby obscuring their original performance practice. Stanko Vraz (1810–1851), conversely, notated only fragmentary melodic outlines without accompanying contextual data. Owing to their lack of accuracy, reflected in rhythmic and metrical inconsistencies, as well as melodically (anomalous features), the few *nineteenth-century* transcriptions preserved in our collection are of limited use for analysing melodic patterns, though they remain valuable as evidence of early documentation practices.

In 1906, the Austrian government funded a major project to collect as much folk song material as possible (including song melodies) from across all the Austrian lands. As part of this initiative, the *Committee for the Collection of Slovene National Songs with Their Melodies* (known as OSNP) was established.[2] The collectors, most of whom were amateurs, were recruited through a public call (OSNP 1907) and provided with guidelines (Štrekelj 1906). These guidelines strongly emphasised that songs were to be transcribed as they were originally sung and that they should not be harmonised. Nevertheless, many collectors, such as Franc Kramar and Ciril Pregelj, continued to harmonise (Golež Kaučič et al. 2007). The collectors were also instructed not to alter the lyrics and song structure, and to avoid transcribing tunes that resembled those already collected. While the latter regulation may have seemed sensible at the time, it ultimately hindered the ability to trace subtle changes in song evolution. Although the project was conceived on systematic foundations, the lack of (audio) recording equipment at the time made it impossible to capture performances accurately, rendering the resulting transcriptions unreliable.[3] The editors of the collection had no access to supplementary material (particularly audio or audiovisual material, which are now considered the most reliable sources) for re-evaluating what was commonly intuitively collected data.

As the project was carried out across many different nations under a single monarchy, it involved a range of socio-cultural, political and methodological compromises. These included decisions regarding what should be collected, considering song popularity (‘What constitutes a folk song?’), genre (work songs, lullabies, folk theatre, etc.), ‘nationality’ or the language of the material (German, Austrian, or Slovenian), and song sorting (lyric, dialects, regions, etc.) (Murko 1929). Many of the melodic transcriptions produced by this project seem simplified and/or inaccurately transcribed. Furthermore, as collectors mostly transcribed individual singers, not groups, they omitted a very rich polyphonic singing tradition (Kovačič 2015). It was not only the collectors’ lack of appropriate skills that contributed to shortcomings in transcription, but also, conversely, their rigid Western music education, which often led them to favour what they considered aesthetically more pleasing tune harmonisations over folk practices. As a result, such transcriptions cannot be regarded as faithful representations of the folk songs of the time (Kovačič 2015; Kumer 1959). Numerous publications of folk songs from the OSNP campaign underwent a series of editorial revisions, but the folk family ballads generally remained unedited.[4]

By the 1920s, folk song researchers had begun to prioritise the collection of both notated melodies and recordings. Štrekelj, in particular, ‘planned extensive and systematic recordings of folk songs and aimed to create an archival sound collection of recordings on wax cylinders’ (Kunej 2022: 40). He also prepared detailed instructions for the recording, archiving, and usage of such material (Kunej 2022). At the time, considerable debate emerged regarding whether recordings should supplement or replace manual symbolic transcriptions. Discussion focused primarily on preserving the ‘authentic’ form of folk songs (Štrekelj 1906), but also addressed the challenges of accurate polyphonic annotations and other transcription issues in music for which staff notation may be inadequate (Kunej 2022). Initially, melodies and lyrics were collected separately; however, as folk song research progressed, these elements were increasingly integrated, along with additional metadata (Murko 1929). While technological advancements made sound recording feasible, the committee responsible for purchasing recording devices initially failed to understand their potential beyond compensating for a lack of notational skills. As a result, Štrekelj was ultimately unsuccessful in persuading the Slovenian committee to purchase audio equipment, and his work remained limited to symbolic transcriptions (Kunej 2005; 2022).

The era of phonograph recordings

In 1914, Matija Murko (1861–1952) managed to acquire the first phonograph in this geographic region to be used for recording folk songs, which Juro Adlešič employed that spring to document performances from White Carniola (sl. Bela Krajina), a region in present-day southeastern Slovenia (Kunej 2022: 45). The appearance of the phonograph brought about ‘a new methodological approach to documenting and studying folk songs’ (Kunej 2022: 46). This shift was further consolidated after 1955, when the Institute of Ethnomusicology ZRC SAZU in Ljubljana adopted its first tape recording device (Kunej 1999). Researchers no longer needed to complete their transcriptions in real time, and informants were no longer required to repeat their songs, as the playback of recordings enabled more accurate transcriptions to be taken from a single performance, rather than relying on approximate live repetitions.

Researchers were now able to focus more closely on performance style, vocal characteristics, precise tempo, non-tempered melodies, and even emotional expression. These advances gave rise to new debates, including those involving Stanko Vurnik (1898–1932), Valens Vodušek (1912–1989), Zmaga Kumer (1924–2008), and Julijan Strajnar (1936) (Kovačič 2015). Through re-listening, re-recording, and converting phonograph cylinders into more accessible formats, researchers gained new insights into the folk songs collected.

Nevertheless, aesthetic, moral, and political influences of the time continued to shape decision-making processes, determining which songs were deemed ‘Slovenian’ enough to be collected. Songs that were considered insufficiently ‘authentic’ were excluded from research. The fusion of the concept of authenticity and cultural nationalism (Bendix 1997) played a central role in the methodology for collecting folk songs in the Slovenian context.

The medium for recording also had a significant influence on the quantity, length, and quality of information, as well as the amount of (acoustic) data captured. For instance, use of a less capable medium, such as wax cylinders, required researchers to favour shorter examples over longer ones, to be selective in choosing what to record, and/or to interrupt performances. Additionally, conversations with the singers were frequently cut out. As the use of recording equipment became more common and fieldwork practices evolved,[5] research participants’ attitudes toward the presence of technology and researchers also shifted, and this in turn affected their performance.

With the evolution of fieldwork and other methodologies, researchers increasingly advocated for a *comparative* approach. Vodušek emphasised the importance of incorporating elements such as context, phonetics, and lyrics. In line with global trends (Lomax 1978; Merriam and Merriam 1964), Vodušek moved beyond mere material preservation for two reasons. First, he viewed music as a dynamic phenomenon, and second, he argued that to define Slovenian folk song, one must compare it to presumed non-Slovenian material (Vodušek 2003). Although several 'musical' issues persisted with the recordings (List 1974), this broader perspective led to changes in material collection, conservation, and methodology.

The digital era

Recording processes eventually developed—not only as concerns capturing of recordings in the field, but also in connection with transcription, reproducibility, and accessibility. From gramophone records to CDs, and from later digital formats such as .mp3 and .wav to digital archives, YouTube, Spotify, and other audio-visual platforms, new technologies prompted re-thinking of conservation practices and research in general. The rise of digital formats necessitated the consistent (re)organisation of physical materials. Consequently, Drago Kunej and Rebeka Kunej developed protocols for collecting (recording and archiving), analysing, and digitising music material (Kunej and Kunej 2020; Stefanija et al. 2022).

Several projects have made folk songs digitally available. These projects have involved radio broadcasts (e.g. the RTV internal archive, the podcast *Slovenska zemlja v pesmi in besedi*), online platforms (such as EtnoFletno and Klik v domovino, the YouTube channel Folk Music Heritage), and CD and digital book editions. These have been joined, most recently, by the Institute of Ethnomusicology's project Etnofon (2023), which is a complete digital collection of the institute's past and present sound publications, all of them having recently undergone digitisation, annotation, and computational analysis. While Strle and Marolt (2012) focused on analysing the conceptual structure and themes in lyrics, Borsan et al. (2023) concentrated on digitising and annotating melodic content and developing an analytical model for pattern discovery. We expand on the latter by explaining the collection's digital representation on the Dezrann platform, updating annotations, and discussing computational analyses applicable to the material. In the following section, we acknowledge similar international projects, discussing their organisation, introducing our dataset's organisation, structure, and contents, as well as demonstrating potential in computational analyses.

Organising

According to Rice, *organising* involves '(1) analysing and describing the data, (2) classifying and categorising the data, (3) labelling categories and classes, and (4) listing categories in some order' (Rice 2017: 30). Approaches to music organisation in the field(s) of (comparative and computational) ethnomusicology tend to be inconsistent (Proutskova et al. 2020). Some are based on Western music organisation and analysis theories (Anagnostopoulou, Giraud, and Poulakis 2013; Conklin and Anagnostopoulou 2011; Ossa 2019; Temperley 2000), while others adapt their analytical theories to the specificities of their research material (Bozkurt 2015; Caro Repetto et al. 2015; Clayton et al. 2022; Nuttall et al. 2019, and others), or work towards a consensus that supports a multicultural comparison of materials (Killick 2020; Lomax 1978; Ozaki et al. 2022; Savage 2020; Wood et al. 2022). This influences the rest of the *theorising* process. The way we describe music, through the classification, categorisation, and labelling of categories and classes, as well as the 'hierarchical' listing of these in a predetermined order, depends on the organisation of observed material (see studies cited above). To establish a digitised music dataset, it is essential to first create a framework for both the collection (of new materials) and the structuring of the data.

Digitised music datasets

Excluding archives that consist of listed works or scans of physical material, we have at present identified more than 15 digitised folk song datasets, some of which are publicly accessible. There are also studies discussing the organisation of folk song material in digital archives (Lidy et al. 2010; Proutskova et al. 2020; Strle and Marolt 2012; Tian et al. 2013) or reviewing existing datasets (Makris, Karydis, and Sioutas 2015). While we will not detail every project, we will briefly introduce the different types of collecting and organising practices associated with digital folk song datasets.

Datasets can be distinguished by contextual diversity, the quantity and quality of (meta)data, the motivation for digitisation, content type (notation, audio, lyrics), and varying levels of online availability. Content can be categorised into datasets that focus on multiple traditions (Bertin-Mahieux et al. 2011; Porter and Serra 2014; Porter, Sordo, and Serra 2013; Sapp 2005; Wood et al. 2022), or those focusing on culturally similar material, such as Greek (Makris, Karydis, and Sioutas 2015; Papaioannou et al. 2022), Indian (Singh et al. 2022; Srinivasamurthy et al. 2021), Dutch (Van Kranenburg, de Bruin, and Volk 2019), Latin (dos Santos and Silla 2015; Silla, Koerich, and Kaestner 2008), Basque (Conklin 2011), and Georgian (Rosenzweig et al. 2020) collections. Some datasets are created for a specific task, such as ethnomusicological research (Porter, Sordo, and Serra 2013; Strle and Marolt 2012; Van Kranenburg, de Bruin, and Volk 2019; Wood et al. 2022), pattern-matching (Borsan et al. 2023; Conklin 2021; Conklin and Anagnostopoulou 2011; Van Kranenburg et al. 2011; Neubarth and Conklin 2016; Ren 2016; Savage et al. 2022, and others), and mood or emotion recognition (Gómez-Cañón et al. 2023; Makris, Karydis, and Sioutas 2015; Pesek et al. 2017). These do not necessarily share common data types, as they can consist of any combination of text, scores, audio, lyrics (where applicable), and metadata. Furthermore, they can be released under a variety of accessibility and reproducibility licenses.

These criteria, along with an understanding of the context and data reliability, influence the future use of digitised materials and play a significant role in scientific discovery and practice. For instance, a dataset with downloadable scores but minimal metadata may be valuable for tasks focusing on music structure, such as melodic pattern matching and machine learning for similarity recognition. However, without adequate metadata and supporting research, this material may not be suitable for ethnomusicological studies beyond the scope of the specific collection project. Conversely, materials with abundant metadata, images, and other computationally unreadable or non-downloadable formats are limited in their ability to be subjected to computer processing.

Data that is not aligned with open-source principles is commonly considered an inadequate contribution to the scientific community (for elaboration on this matter, see Weigl et al. 2019; Weigl et al. 2021; Wilkinson et al. 2016). Therefore, the theoretical and material foundations for collecting and organising data are crucial for (computational) music research possibilities.

Structuring

Rice defines *structuring* as the creation of two classifications, *typologies* and *taxonomies*, and as *data modelling* (Rice 2017: 31). *Typologies*, which are primarily conceptual and serve as a systematic foundation for comparisons, consist of constructed categories used to describe particular objects in opposition to others. Examples include melodic contours (Adams 1976; Anagnostopoulou, Giraud, and Poulakis 2013; Cornelissen, Zuidema, and Burgoyne 2021; Huron 1996), cadences versus non-cadences (Van Kranenburg and Karsdorp 2014), and pattern versus antipattern (Conklin 2013). *Taxonomies*, by contrast, are broader quantitative classifications which, based on measurable characteristics, sort individual objects into general categories and define their (hierarchical or relational) position in relation to objects with which they form homogenous groups. In music, examples include the Hornbostel-Sachs musical instrument classification (Lee 2019; von Hornbostel and Sachs 1914) and various folk music genre classifications (Neubarth et al. 2012; Ren 2016; Savage 2020). Lastly, *models* deal with the probability of events, highlight the relationships among observed data, and act as 'heuristic tools for asking questions about particular research projects' (Rice 2017: 32).

(An)notation systems

Meticulous organisation, systematic labelling, and effective archiving practices are not merely supplementary tasks but core requirements for work with digital datasets. Regardless of the content type, be it music collections, literary works, or any other form of data, these principles underpin accessibility, preservation, and usability.[6] When materials are no longer intended for standalone storage but are expected to undergo (computational) processing, an additional set of protocols is necessary. An example of this is the OWL (ontology web language) (Antoniou and van Harmelen 2004; McGuinness and van Harmelen 2004) and its implementation of the CIDOC Conceptual Reference Model. 'The OWL representation facilitates encoding and reasoning over a genre ontology, while the CIDOC model enables a representation of complex spatial containment and proximity relations among geographic regions' (Urkizu et al. 2012: n. p.), allowing researchers to navigate various types of cultural heritage information or study different toponyms, and to systematically arrange content based on its 'complexity'. The application of music ontology within the semantic web has been addressed by several authors (Proutskova et al. 2020; Raimond et al. 2007, and others). While broad ontological categories might suffice for any music data, more detailed modelling tailored to specific cultures and use cases is often required. This, however, inevitably leads to (partial) data incompatibility across studies, even when shared reference models (CIDOC, FRBR, etc.) are employed.

Images and texts share a common characteristic—they are represented in specific forms, such as pixels for images and alphanumeric symbols for the written word. Unlike some standardised representations, the realm of music has no optimal universal representation. This issue has led to the development of formats such as musicXML (or .sib, or .finale), MEI (Rizo and Marsden 2019), MIDI (Eerola and Toiviainen 2004; Huang et al. 2013; Rothstein 1992), ABC,[7] LilyPond (Nienhuys and Nieuwenhuizen 2003), and GUIDO music notation (Hoos et al. 2001; Renz 2002), among others. The varied representations of music significantly influence the complexity of data modelling and analysis tasks, and they profoundly shape the outcomes of research. Ideally, efficient computational structuring methods can allow researchers to skip Rice's initial steps and move directly to data explanation and synthesis; but this is only feasible when the model aligns with the research objectives and its structure and function is comprehensible, which is a rare occurrence.

Computational ethnomusicology: general ideas

Substantial efforts have been made to uncover effective (computational) research practices for genres other than Western (art) music; some of these research practices trace back to the pre-computational era. Notable contributions include the work of Samuel Bayard (1950), James R. Cowdery (1984), Kurt Blaukopf (1993; see also Zembylas 2012), Isabelle Mills (1974), Alan Lomax and colleagues (1978; see also Savage 2018; Wood et al. 2022), Alan P. Merriam (Merriam and Merriam 1964), Mantle Hood (1982), and Rice (2017). These foundational works continue to serve as a source of inspiration for contemporary (computational) ethnomusicology, especially in the areas of tune-matching and other pattern-matching tasks, as well as in music element definition and labelling.

Bayard's theoretical framework addressed the nuances of tune development, the relationship between recording and origin time, and the selection of descriptors for understanding music patterns, stressing the subjectivity of observation (Bayard 1950). He discussed tonal range, rhythm, melodic progression, the order of stressed notes and, notably, defined the concept of a *tune family*,[8] which is highly relevant to computational music research.

Cowdery noted that, '[a]lthough Bayard laid much important groundwork, he did not provide sufficient models for these potential applications' (Cowdery 1984: 495). Instead of combining descriptors for the tune as a whole, Cowdery proposed three principles for working with folk song material: the *outlining*, *conjoining*, and *recombining* principles. The first principle 'allows us to compare wholes to wholes, and the second provides for comparing sections to sections' (Cowdery 1984: 498). The third principle combines the first two by viewing melodic combinations as a system of potentialities, rather than a fixed sequence of events. This approach demonstrates that 'motives can recombine in various ways, expanding or contracting, to make new melodies which still conform to the traditional sound' (Cowdery 1984: 499).

Both of Cowdery's contributions have been acknowledged in computational studies, particularly in folk music *pattern matching* and *sequence alignment* (Boot, Volk, and de Haas 2016; Borsan et al. 2023; Bountouridis et al. 2017; Pendlebury 2020; Ren et al. 2018; Savage et al. 2022; Savage and Atkinson 2015; Van Kranenburg, Volk, and

Wiering 2013; Volk et al. 2007; Volk and Van Kranenburg 2012; Volk, de Haas, and Van Kranenburg 2012).

Computational work with data also necessitates a comprehensive, systematic breakdown of what and how we annotate and analyse, as exemplified by Ossa's theoretical framework for defining basic music parameters in folk song analysis (Ossa 2019). He discusses the analysis of song, melodic, rhythmic and lyric structure, scale types, ranges, and other music descriptor categories. Lomax's concept of *cantometrics* (Lomax 1978), later elaborated on by Wood et al. (2022), presents an alternative method for encoding musical elements, establishing a numerical encoding system that can identify patterns in singing styles and musical structures across cultures, with the aim of understanding cultural and social dynamics. It correlates music and cultural factors, offering insights into human behaviour and societal relationships. These principles align closely with the approaches of many ethnomusicologists of the time, including the aforementioned Vodušek.

Alongside notational and pattern-based approaches, there is a growing strand of digital ethnomusicological research that focuses on comparative, cross-cultural analyses of diverse materials such as speech, singing, and instrumental sound. Often drawing on evolutionary and behavioural frameworks, these studies make use of large-scale, annotated datasets to address long-standing questions regarding music's origins, functions, and universality. For instance, Mehr et al.'s 'Universality and Diversity in Human Song' (2019) presents a 'natural history of song' by combining ethnographic texts and audio recordings from a globally representative sample of societies. Employing computational social science methods, it demonstrates that music can be considered a human universal, consistently linked to specific behavioural contexts (e.g. infant care, healing, love), and it shows that musical features—such as pitch, tempo, and rhythmic complexity—systematically align with social functions. In a similar vein, Ozaki and Savage's 'Many Voices' project (Ozaki et al. 2024) explores the relationship between music and language by analysing matched recordings of song, speech, recited lyrics, and instrumental music across more than 50 languages. Their findings suggest a 'musi-linguistic continuum', emphasising both shared and distinct acoustic characteristics across expressive modalities. Collectively, these studies illustrate how computational tools are increasingly applied to rich humanistic data, enabling investigations into profound questions about cultural diversity, commonality, and the cognitive underpinnings of music.

In contrast to these all-encompassing systems, David Huron, among others, focuses on a single descriptor, such as *melodic contour* (Huron 1996). This work emphasises the perceptual relevance and importance of 'melody direction' over absolute pitch values. It identifies nine common melodic archetypes for phrases, based on the relationships between initial and final pitch values and a median of middle pitch values. These relationships, Huron argues, provide a more perceptually relevant description of melodic character than absolute note values.

Ultimately, in Cowdery's words,

> [a] traditional musician will not evaluate a new tune or version by comparing it to some faceless archetype […]. If we wish to understand this process we must look for *principles* —the overlapping and flexible ways in which musicians work with their materials—rather than looking for categories to impose from outside. (Cowdery 1984: 498)

This is a task that has yet to be fully mastered in computational analyses. However, what is already achievable is the incorporation of some contextual meaning. As Pendlebury stresses,

> [t]he examination of tunes in the contexts of their source documents elucidates the cultural factors that influenced their reuse over history, such as the development of print technology, military campaigns, trends in commercial theatre, and the mass production and use of instruments […]. (Pendlebury 2020: 92)

It is only when we understand the materials collected—their organising and structuring—that we can begin the process of explanation and synthesis. Until then, these encodings and the algorithms comparing the encoded materials are, akin to Floridi's concept of genes, 'a type of predicative and effective/procedural information […] [which as] dynamic procedural structures […] together with other indispensable environmental factors, contribute to control and guide the development of organisms' (Floridi 2010: 75). Thus, biological information 'is information *for* something, not [yet] *about* something' (Floridi 2010: 76). Therefore, following the description of collection practices, the next section explains how our dataset is organised and structured.

The dataset of Slovenian folk song ballads: family ballads

Slovenian researchers participated in the international framework of European scholars studying the ballad tradition, an initiative that in 1966, adopted a shared definition of ballads in the European context: 'a ballad is a song that tells a story with dramatic emphasis' (Kumer 1998: 31). In summarising the numerous definitions and genre classifications present in international folklore studies, Golež Kaučič emphasises that the 'ballad is defined through genre and tradition' (Golež Kaučič 2018: 350) as well as through oral transmission. According to Golež Kaučič, ballads recount the story of an event and its outcome in a dramatic manner. They are also embedded within a dynamic process of variant creation, maintain an evolving relationship with their context, and serve a specific function within an individual community (Golež Kaučič 2018: 351).

Our collection comprises 402 monophonic transcriptions of Slovenian folk song ballads, previously transposed to G-centred tonality by the curators of the physical collection for analytical purposes. These transcriptions, which were made in different periods (see section on 'Collection and Conservation Practices' above) were retrieved from the archives of the Institute of Ethnomusicology at the Research Centre of the Slovenian Academy of Sciences and Arts. The scores, lyrics, and available recordings associated with this study have been published as an open-access dataset (Borsan et al. 2024). At present, only the first verse of each ballad is published. However, it is important to note that ballads are never limited to a single verse. Each melody-verse pair was previously categorised according to the content of its lyrics and contextualised with metadata. In the current dataset release version, these entries are further enriched with additional annotations, which are divided into two groups: those describing the song as a whole and those describing individual phrases within each song.

Organising the dataset

Metadata

These songs have previously been organised into 36 distinct topic types, each with between one and as many as 103 variants. They originate from 22 different regions, with the

Table 1. Most frequent song types, regions, transcription year, and singer type, with the respective numbers of categories and occurrences.

Song Types (35)	Regions (22)	Transcription Year (67)	Singer Type (6)
286 (103)	Styria (111)	1957 (37)	F_Solo (292)
252 (82)	Upper Carniola (86)	1970 (31)	M_Solo (39)
287 (45)	Lower Carniola (52)	1908 (21)	Unknown (52)
256 (44)	Raba (HU) (35)	1907 (16)	F_Group (14)
267 (28)	Prekmurje (34)	1961 (16)	MIX_Group (5)
254 (11)	Littoral (24)	1958 (14)	M_Group (2)
248 (10)	Inner Carniola (22)	1963 (12)	–
258 (9)	White Carniola (10)	1910 (12)	–
277 (9)	Carinthia (AT) (8)	1839 (11)	–
279 (9)	Littoral (IT) (6)	1982 (11)	–

majority coming from Styria, Upper Carniola, and Lower Carniola (Table 1 and Figure 2). The corpus spans a period of 176 years (1819–1995), comprising transcriptions and recordings with identifiable dates from 68 distinct years. The earliest transcription dates to 1819, while the most recent was made in 1995 (Table 1 and Figure 3). Each type has been assigned a topic title; however, the context of the individual song was further determined by the opening line of its first verse.

Most songs in our collection were performed by solo female singers (Table 1 and Figure 4), with those for solo male singers the next most common. Klobčar (2014)

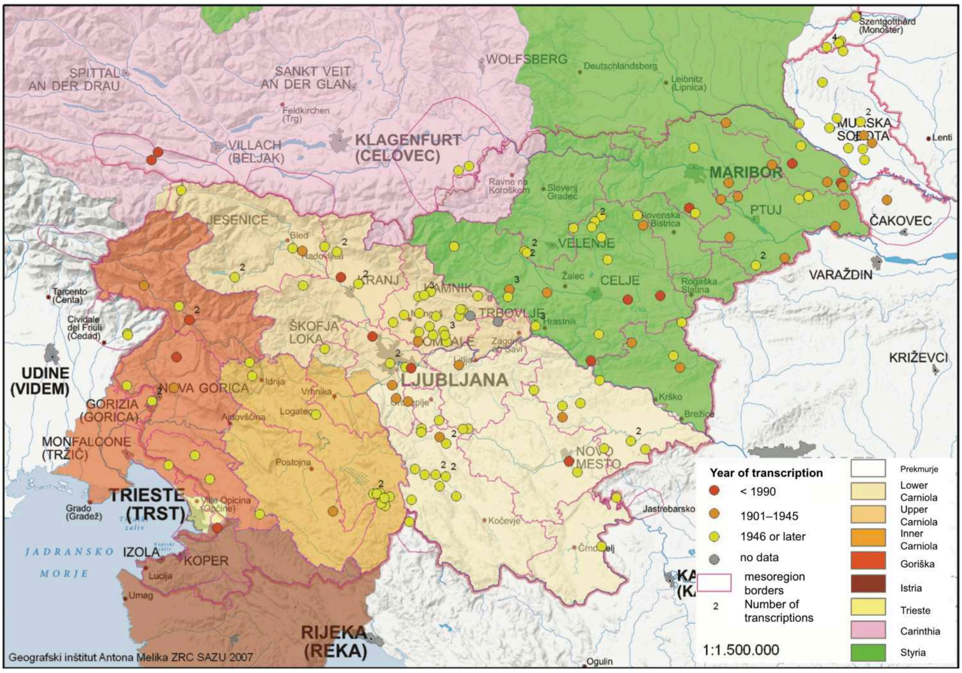


Figure 2. An example from *Slovenske ljudske pesmi 5* (Golež Kaučič et al. 2007: 909), illustrating the occurrence of the song type *Nevesta detomorilka / Infanticide Bride* across various Slovenian regions, along with approximate dates of appearance. The legend is translated by the authors; the original version along with many other maps is available in Golež Kaučič et al. (2007).

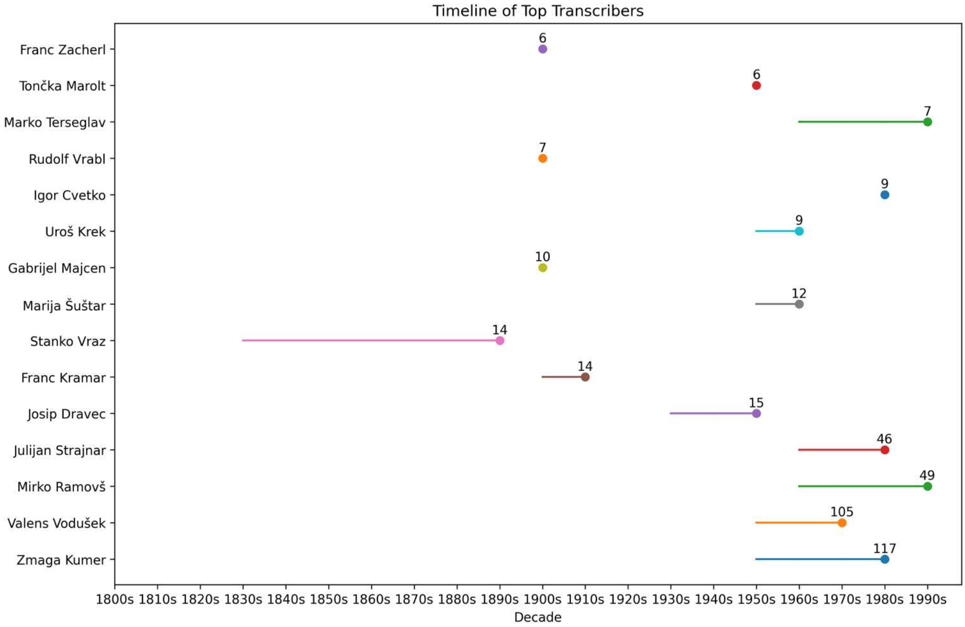


Figure 3. The most represented transcribers, with the number of songs transcribed and the years of transcriptions.

discusses the prevalence of female singers and their role as the principal bearers of the ballad tradition, suggesting that these phenomena are the results of shifts when songs lose their original social impact and consequently attract less male interest. Klobčar's study also explores the link between women singers and these songs in the context of identity-construction in the nation-building process, which may be relevant to our material as well. As our collection primarily consists of monophonic tunes, singing in groups is notably underrepresented. A consequence of the methodology used for collecting the material is that the ballad songs recorded are mostly monophonic melodies.

Researchers and collectors adopted an exploratory method for gathering data, engaging directly with individual singers and, on occasion, capturing songs during impromptu group singing sessions, which often featured multipart harmonies. Therefore, although most of the songs in the collection are monophonic, this does not necessarily mean that they were sung in unison in spontaneous situations; in fact, such cases were rare. Given this and because a computerised comparison of the material, and consequently the analysis, are easier with consistent material, the analysis of musical features presented here will also be based on monophonic melodies.

Corpus availability/interaction

The dataset (including scores in .musicXML, recordings in .mp3, metadata, and annotations in .csv/.json files) is released under the open CC BY-NC-SA 4.0 license and available for download at http://algomus.fr/data. The songs are also available on the open-source platform Dezrann,[9] where all annotations for each song are visualised and, when audio files are available, synchronised with the scores. The Dezrann interface is available in seven languages, including Slovenian and English; however, most of the corpus data

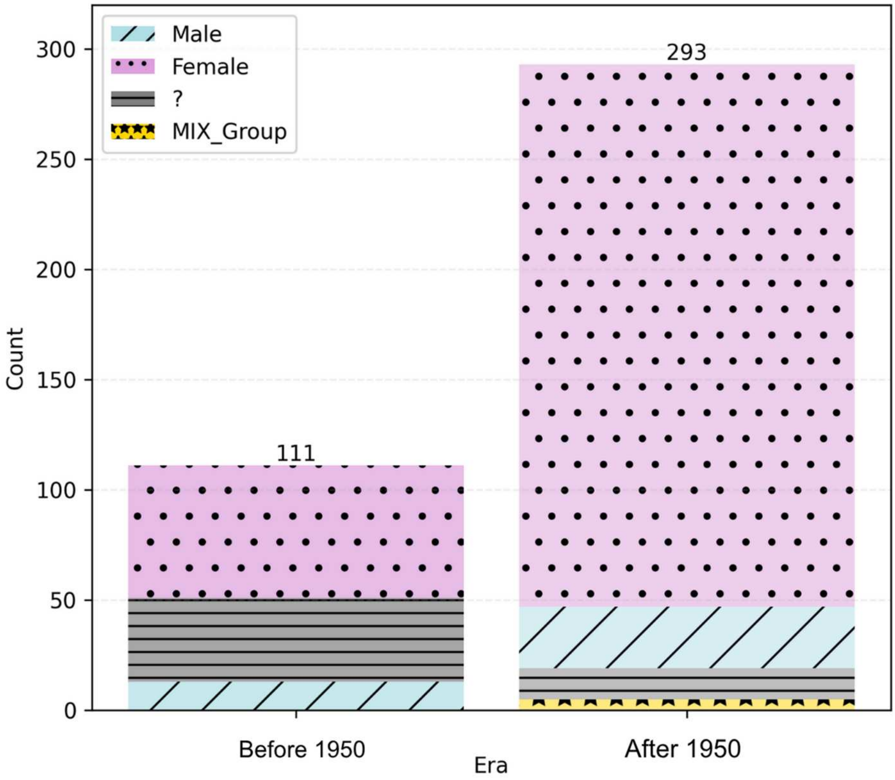


Figure 4. The count of male, female, mixed group and unknown singers in the eras before/after 1950.

itself is in Slovenian and has subsequently been translated into English.[10] The Dezrann visualisations are set to default to the initial preassigned labels, which can be altered, deleted or added by the user. If the original sources indicate minor melodic variations occurring after the first verse, for now these are excluded, but they can be found in other editions, such as Golež Kaučič et al. (2007).

Structuring the dataset

In our dataset structure, we classify descriptors into three categories: non-music descriptors, lyric descriptors, and music descriptors. The internal structure of the first category draws on the collections and archives that collected and curated the sources. The naming of some of the *non-music* descriptors or *metadata* was adapted to fit the computational system, with each one also translated into English (Table 2). The *lyric* descriptors were also partially derived from the systems used in Slovenian folk song collection and were further refined by some of the authors of this paper (Table 3). These include features such as the first verse, as well as metric and rhyme verse structure. Lastly, the *music* descriptors are informed by the ideas presented in the 'Structuring' section and include all elements listed in Table 4 (such as time signature, phrase number and label, and contour). An example of a song with some of the labels is displayed in

Table 2. Non-music descriptors or metadata of the released dataset.

Descriptor	Categories/Explanation	Example
ID	Derives from the printed collection and follows the sequence: type, variant, additional marker	*256.2.A2*
Type	One type title per song	*Mačeha in sirota / The Stepmother And Her Stepchild*
Variant	The first verse of the song	*Pšenička na polju že zori / Wheat in the field is already ripening*
Region	Regional information on where the song was transcribed or recorded	*Dolenjska / Lower Carniola*
Transcriber / Collector	The person who collected or transcribed the song	*GNI (Zmaga Kumer)*
Year	Year of transcription/recording	*1963*
Singer(s)	The person who performed the song	*Frančiška Lavrič*
Singer(s) Type	Female or male solo, female or male group, mixed group	*F_Solo*

Table 3. Lyric descriptors.

Descriptor	Categories/Explanation	Example
ID	See Table 1.	*256.2.A2*
Verse Structure (meter)	A number or set of numbers, indicating the number of syllables in a phrase	*6*
Verse Structure (rhyme)	Alphabetic character (starting from M), indicating the resemblance of rhyme of the final word in a phrase, plus two types of refrain—full (R), where the totality of a verse line is a refrain, or half (r), where a part of a verse line is a refrain. In both cases, they are generally composed of non-lexical vocables, such as 'tralala' (286.23), or 'jupajde' (156.16.A3), or several repetitions of the same word (273.7).	*MNOPOP*

Table 4. Music descriptors.

Descriptor	Categories/Explanation	Example
ID	See Table 1.	*256.2.A2*
Time signature	Time signature(s) of the song in order of first appearance.	*6/8*
Upbeat	Upbeat expressed in quarter notes.	*2*
Note count	/	*37*
Measure count	/	*9*
Range	Melodic range as an interval and in semitones.	*m9 (13)*
Pitch mean	Melodic pitch mean of a song.	*65.94*
Pitch direction	Describing ascending, descending, and equal relationship between the first and the final pitch.	*Ascending*
Tone set	A number and set of unique pitch classes in a song.	*7: G, A, B, C, D, E, F#*
Leading Tone	Presence/absence of the note f# (in respect to notations transposed to G).	*YES*
Phrase number	Enumerated melodic phrases, in order of appearance.	*6*
Phrase labels	Alphabetic characters indicating melodic relationships among phrases. The same letter indicates melody that is identical, transposed, or partially adapted (to lyrics, range, singing mistakes with repetition, etc.). All major differences are marked with a new letter.	*AABABA*
Contours	The melodic shape of individual phrases in a tune.	↗, ↘↗, ↗, ↘↗, ↗, ↘↗

Figure 5, while a detailed explanation of the music descriptors can be found in the README[11] file accompanying the dataset, as well as in the following section.

Explanation and synthesis

In this section, we present statistical information on music descriptors, including scale, range, intervals, and structural labels for both lyrics and melody. We also focus on

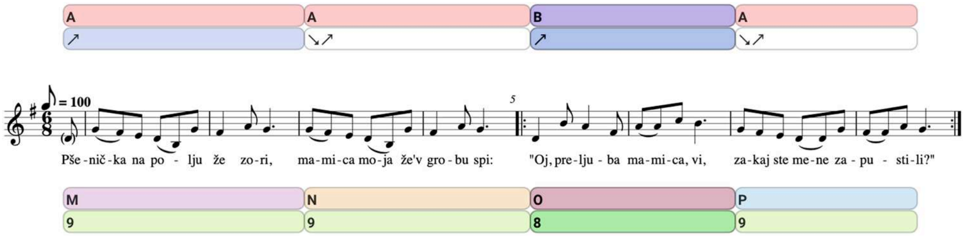

A
↗
A
↘↗
B
↗
A
↘↗
= 100
5
Pše-nič-ka na po - lju že zo-ri, ma-mi-ca mo-ja že'v gro-bu spi: "Oj, pre-lju - ba ma-mi-ca, vi, za-kaj ste me-ne za - pu - sti-li?"
M
9
N
9
O
8
P
9

Figure 5. *Pšenička na polju že zori* (type 256.A2, variant 2, see Tables 3–5) with labels (from top to bottom) of melodic phrase structure (AABABA), contours (↗, ↘↗, ↗, ↘↗, ↗, ↘↗), and rhyme (MNOPOP) and syllable (998989) verse structures. This example is a screenshot of an interactive score and visualisation of annotations on Dezrann.

specific correlations between metadata, music descriptors, and melodic patterns, and provide a synthesis of the results.

The songs are typically sung in 2/4, 3/4, 4/4, or a combination of the first two metres. In terms of range, they usually fall somewhere between a minor seventh and a perfect octave. Extremely narrow melodies, covering a range of less than a perfect fifth, are exceptionally rare. The songs tend to conclude on a note higher than the starting note, rather than on the same or a lower note; most end on a third (B), first (G) or fifth (D) scale degree. There are only ten exceptions, where melodies do not resolve on the notes of the tonic chord, and these are often due to errors such as incorrect notation (for instance, Vraz's melody sketches such as 248.1; Figure 6(a)), lack of transposition to G-major (259.3; Figure 6(b)), or inaccuracies in singing as indicated in the collector's notes (257.6, 286.100). Another possibility is that the song follows an older an older modal scale, of a kind prevalent in the periphery of present-day Slovenia, which lies outside the major–minor melodic system (283.6, 248.20, etc.; Figure 6(c and d)).

Most melodies are composed of approximately six or seven different pitch classes; however, examples of 'few-tone' melodies (pentatonic, tetratonic, and tritonic) are also present. Among the 25 different combinations of pitch classes observed, more than half are derived from the (G-)major scale or the (G-)major scale without the third degree (Table 5).

There are some descriptors that are worth observing in direct relation to phrases (Tables 6 and 7). There are seven different possible phrase structures, comprising of a range from two to eight phrases, with the most common being a structure of four phrases. These four-phrase structures are mostly composed of contrasting melodic material, with the ABCD pattern being the most frequently represented, while AB and ABAB structures rank as the second and third most prevalent overall. It should be noted that this is a generalised structure and that there are often inconsistencies across verses in a song. The most common rhyme type is MNMN, followed by MNOP (four contrasting parts) and MN (two contrasting parts).

Year-related correlations

We examined the correlations between diverse musical descriptors and the year of transcription. While previous research has explored how tone sets, range, and length may change with the song's age, our study is constrained by the available data: the date of transcription does not necessarily reflect the age or period of origin of the song itself.

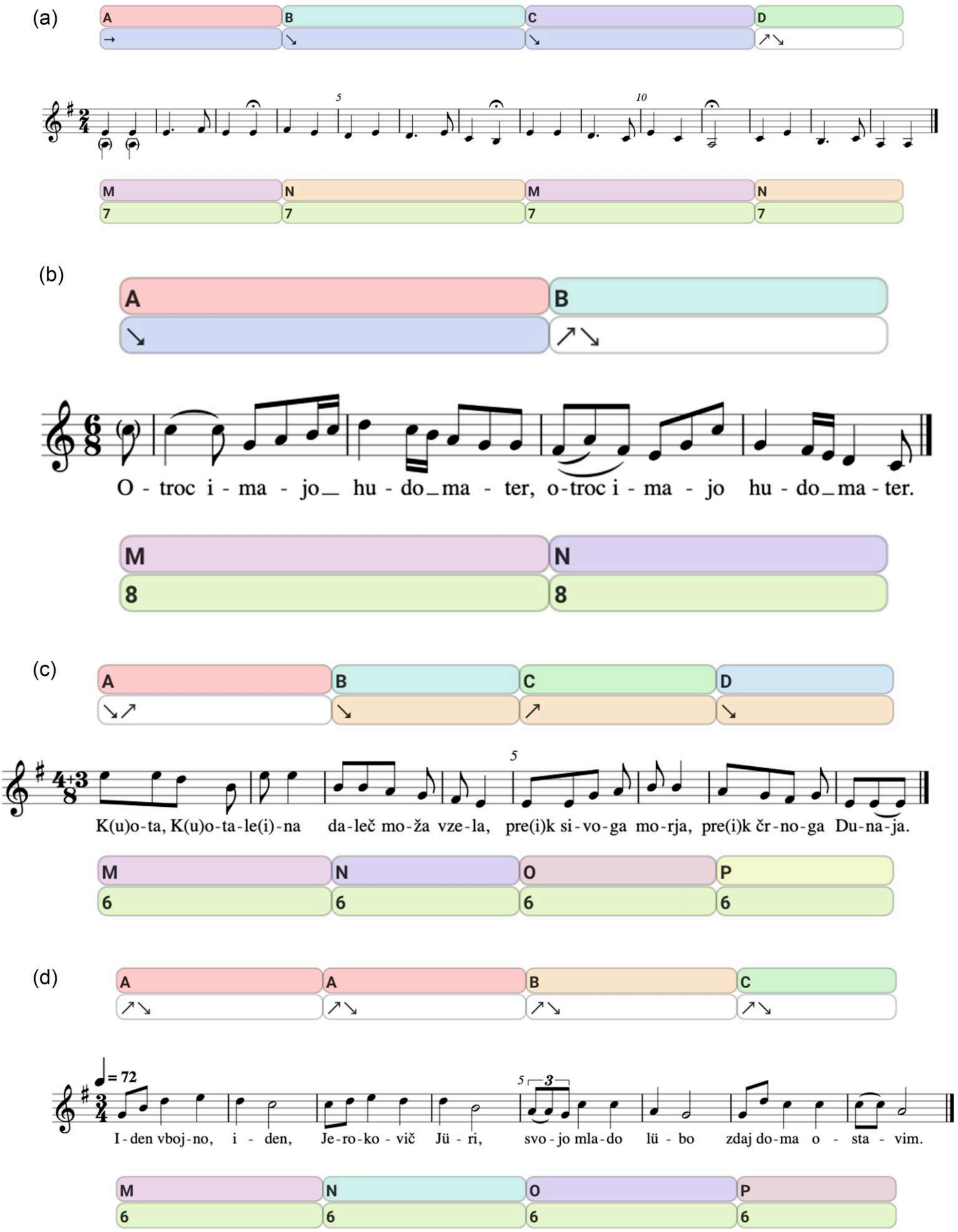

(a)
A
→
B
↘
C
↘
D
↗↘
5
10
M
7
N
7
M
7
N
7
(b)
A
↘
B
↗↘
O - troc i - ma - jo_ hu - do_ma - ter, o-troc i - ma - jo hu - do_ma - ter.
M
8
N
8
(c)
A
↘↗
B
↘
C
↗
D
↘
5
K(u)o-ta, K(u)o-ta-le(i)-na da-leč mo-ža vze-la, pre(i)k si-vo-ga mo-rja, pre(i)k čr-no-ga Du-na-ja.
M
6
N
6
O
6
P
6
(d)
A
↗↘
A
↗↘
B
↗↘
C
↗↘
♩= 72
5
3
I-den vboj-no, i - den, Je-ro-ko - vič Jü - ri, svo - jo mla-do lü - bo zdaj do-ma o - sta - vim.
M
6
N
6
O
6
P
6

Figure 6. Song examples from Dezrann with resolution 'abnormalities' in respect to 'tonic' (G): (a) 248.1 (approximate sketch by Stanko Vraz), (b) 259.3 (not transposed), and (c) 283.6 and (d) 248.20 (not in major-minor modes) (**a**) 245.1 (**b**) 259.3 (**c**) 283.6 (**d**) 248.2.

Consequently, we cannot determine when particular melodic variants first emerged. Since only the year of transcription is known, our findings provide an overview of descriptors across two broad periods: 'older' (before 1950) and 'newer' (1950). The majority of transcriptions from the older period were collected in Styria, while most transcriptions from the newer period came from Upper Carniola (Figure 7). It is important to acknowledge

Table 5. Most frequent music descriptors, with the number of categories and occurrences.

Time Signatures (35)	Range (16)	Pitch Direction (3)	Number of Pitch Classes (7)	Name of Pitch Classes (25)
2/4 (91)	P8 (101)	Ascending (179)	6 (182)	G, A, B, C, D, E, F# (144)
3/4 (87)	m7 (100)	Equal (116)	7 (148)	G, A, B, C, D, F# (92)
4/4 (67)	M6 (66)	Descending (108)	5 (65)	A, B, C, D, E, F# (41)
2/4, 3/4 (46)	M9 (32)	–	4 (4)	G, A, B, C, D, E (36)
3/4, 4/4 (22)	P5 (27)	–	8 (2)	G, A, B, C, D (20)
3/8 (21)	m9 (21)	–	3 (1)	A, B, C, D, E (20)
6/8 (17)	m6 (20)	–	2 (1)	G, A, B, D, E, F# (7)
2/4, 3/4, 4/4 (12)	M7 (13)	–	–	G, B, C, D, E, F# (6)
7/8, 3/4 (3)	M10 (6)	–	–	B, C, D, E, F# (6)
2/4, 4/4 (3)	m10 (6)	–	–	G, A, B, D, F# (5)

Table 6. Up to ten most frequent *melodic phrases and lyric* elements in descriptor categories and the number of times they occur.

Number of Phrases (6)	Phrase Labels – Melody (40)	Lyric Structure – Syllables Per Verse (15)	Phrase Labels – Verse (33)
4 (257)	ABCD (134)	8-7 (236)	MNMN (101)
2 (83)	AB (71)	6 (49)	MNOP (80)
6 (29)	ABAB (38)	7 (41)	MN (76)
3 (22)	AABC (21)	6-5 (29)	? (41)
5 (6)	ABCB (13)	8 (24)	MNNO (27)
8 (6)	ABBC (12)	10-9 (9)	MNOPOP (20)
–	AABA (11)	10 (8)	Unsegmented text (11)
–	AA (11)	Heterometric (3)	MNN (7)
–	ABCDCD (10)	9 (2)	MNRN (5)
–	ABC (9)	? (2)	MNO (4)

Table 7. Up to ten most frequent melodic contours (overall, first phrase, last phrase) and the number of times they occur. For explanation of abbreviations, refer to README file.[a]

Contour Combinations (237)	Contour: First Phrase (9)	Contour: Last Phrase (9)
↗↘, ↗↘ (15)	↗ (120)	↗↘ (181)
↗, ↗→, ↗↘, ↗↘ (12)	↗↘ (95)	↘ (130)
↗, ↗↘ (12)	↘ (62)	→↘ (42)
↗, ↗↘, ↗, ↗↘ (11)	↘↗ (57)	↘↗ (17)
↘↗, ↗↘, ↗↘, ↘ (10)	↗→ (26)	↗→ (13)
↘, ↘ (10)	→↘ (15)	↗ (9)
↗↘, ↘ (6)	→↗ (12)	↘→ (8)
↗, ↗, ↗→, ↗↘ (6)	→ (11)	→↗ (2)
↗↘, ↗↘, ↗↘, ↗↘ (6)	↘→ (5)	→ (1)
↗, ↗→, ↗→, ↗↘ (6)	–	–

[a] The file can be accessed through https://algomus.fr/data.

that a singer in the newer period might have performed songs from both temporal categories and that many similar variants were collected from certain regions, while others remain underrepresented. This adds complexity to our analysis, making it difficult to distinguish songs solely based on the year provided.

Transcriber-related correlations

When discussing the two eras, it is also important to consider the stylistic tendencies or biases of transcribers (see the ‘Collection and Conservation Practices’ section

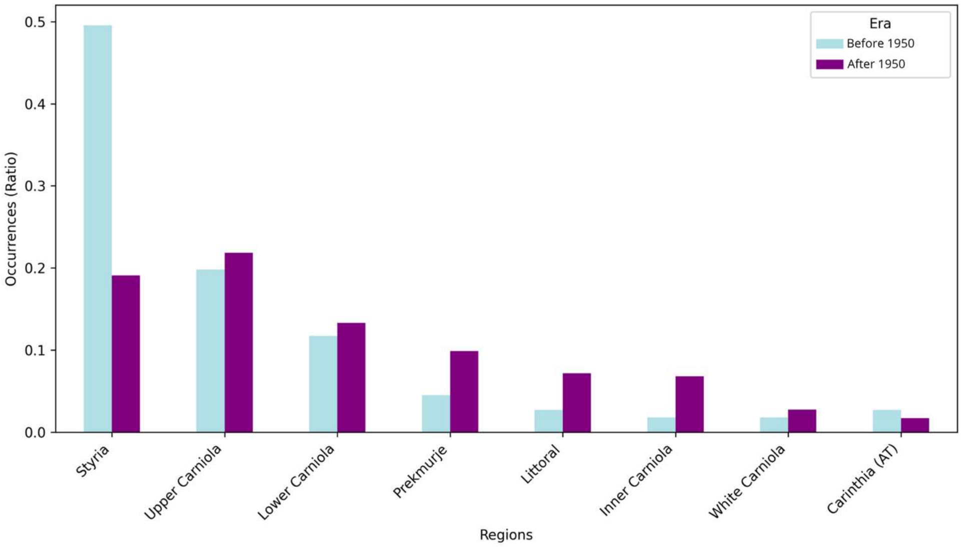


Figure 7. Songs before/after 1950 by region (not all regions are shown).

above); after all, what we are comparing are transcriptions made in a given year, not the year of origin of the songs themselves. We took the most common transcribers before and after 1950 (Figure 3)[12] and compared several descriptors. While most of these are equally (un)common in the two groups or cannot be examined further due to unbalanced data (45 for older versus 262 examples for newer transcriptions), the smallest tone set and the metric verse structure of eight syllables appear almost exclusively in the older examples (Figure 8). This finding confirms some of Vodušek's

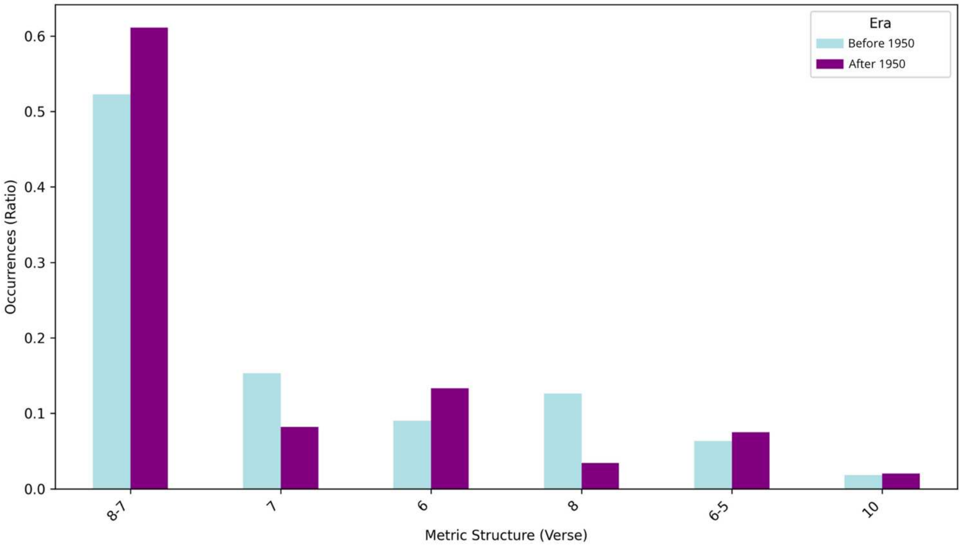


Figure 8. Most frequent (more than 1%) metric structure of lyrics before/after 1950.

manual observations (Vodušek 2003). Older tunes were also more commonly notated in 6/8 time signature,[13] had a wider range (m9, m10), a larger number of phrases (eight), and rarely consisted of more than three contrasting melodic parts (ABCB, ABC, ABAB). As with 'year-related' correlations, we are confronted with the unreliability of older transcriptions, unbalanced data, and the fact that some tunes have many very similar variations from the same era. Further discussion of these issues is provided below (see 'Music Descriptor Analysis').

Region-related correlations

We can retrieve more information about the differentiation among the most represented regions. However, once again, in some regions many variants of the same song have been collected, whereas in others, only one or two were recorded, leaving us with unbalanced data. For the purposes of analysis, we include five of the most frequently represented regions (Table 1) in a brief descriptor analysis.

Music descriptor analysis

We found the *metric structure* of verses to be one of the most common descriptors to appear when differentiating between eras, regions or types. This is expected given that, firstly, melodies tend to adapt to verse (with the reverse occurring less commonly); secondly, a single type shares most of the first verse content; and thirdly, some types tend to be more popular in some regions.

Mixed metric structures (for example, verses with alternating syllabic patterns such as 8–7, 6–5, or 10–9) account for 69% of the corpus, compared to 31% featuring simpler, uniform structures. As anticipated by Vodušek, in correlation with the *era of transcription*, newer songs more frequently exhibit mixed metric structures compared to older

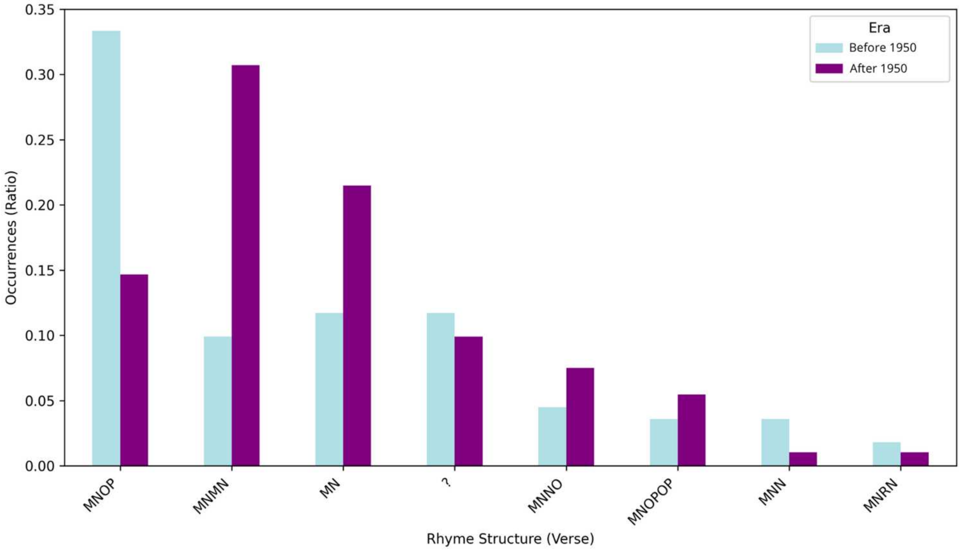


Figure 9. Most frequent (more than 1%) rhyme types in songs before/after 1950.

ones, where simpler structure is more common (Figure 8). The opposite is true for *rhyme*, where older cases more often correspond with the MNOP pattern, while newer songs tend to display repetitive structures such as MNMN, MN, or MNNO (Figure 9). No region was particularly notable in the analysis of lyric structure (for further discussion, see Terseglav 1987).

Similar to lyric metric structures, mixed *time signatures* are more commonly found in newer transcriptions than in older ones (Figure 10). The older transcriptions are generally transcribed in either 2/4, 3/4, or 4/4, or in a combination of two of these time signatures. The 6/4 or 5/4 signatures are only found in older songs. We found that songs in 2/4 m are most prevalent in Raba, where they represent about 80% of the regional corpus, followed by Prekmurje (∼45%) and Styria (∼30%). The most common metre for Upper Carniola is 3/4 (∼35% of songs from this region), whereas for Lower Carniola it is 4/4 (∼38%). All these findings about structures are highly dependent on how collection was carried out. As discussed in the section on 'Collection and Conservation Practices' above, the more contemporary the transcription, the more numerous (and detailed) the data. The prevalence of mixed time signatures in older material could be a result of, for example, transcription style, knowledge, and manners of the transcriber rather than the actual age of the song.

However, this observation does not hold for *contours* or melodic arches, which offer a coarse representation of melodic sequences and do not rely on the absolute accuracy of the transcription. We also found that contours are more significantly influenced by phrase position than other descriptors. For instance, the most prevalent contour is convex (down-up-down), and it is found more frequently in the last phrase than the first one, contrasting with other contour types (Table 7). Notably, descending arches are more common in the last phrase, while ascending arches predominantly occur in the first phrase. While it is difficult to ascertain the overall relevance of these results, the trend is clear: first phrases tend generally to go *up*, while last phrases go *down* (Huron 1996). For the middle phrases, the results are more mixed (Borsan et al.

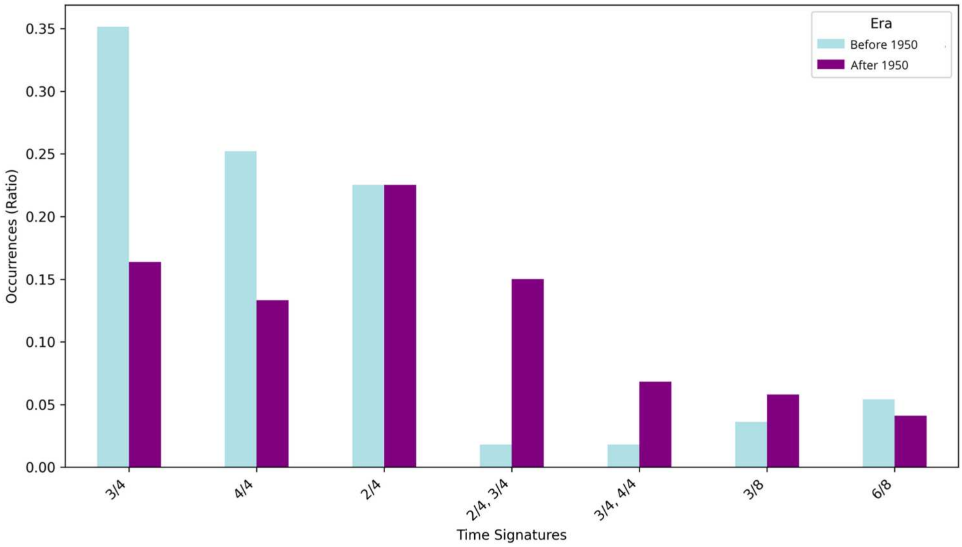


Figure 10. The most frequent time signatures (more than 1%) in songs before 1950.

2023). Moreover, the results for the last phrase are valid for all five regions (the descending or convex contour is present in over 50% of cases in every region), but they are not consistent for the first. Specifically, an ascending contour is most represented in Upper (~50% of songs from this region) and Lower Carniola (~40%); in Raba, concave (~50%) and ascending (~20%) contours are the most frequently occurring; in Prekmurje, convex (~40%) and concave (~20%) contours are prevalent; and in Styria, interestingly, both descending (~30%) and ascending (~30%) contours predominate. We found no direct connection between contours and the *era* of transcription.

The correlations between *year* and *tone set* across the entire corpus did not reveal any substantial differences. However, with regard to *region*, melodies employing a full heptatonic ('major/minor'-type) scale are most common in Raba and Prekmurje (~60%), while in the remaining three regions, such full scales are present in ~40% of cases. The smallest tone sets (consisting of three and two tones) are found in Prekmurje, whereas the largest (consisting of eight tones) occur in Lower Carniola. Most songs in all regions include six or seven pitch classes.

Melodic sequence analysis

Melodic phrases typically consist of approximately eight tones, with each pair of subsequent tones a minor second apart and within a range of a perfect fifth. The majority of melodies are syllabic. To explore melodic relationships between phrases across types and variants we tested[14] the most frequent melodic sequences from the two most common types, 252 and 286 (Borsan et al. 2023).

Sub-sequences consisting of the first and last four tones of a phrase rarely occur in other types or in the same position of the same phrase, and they do not often correspond to the most frequent combinations. This varies across songs. For instance, the sequence f#–a–d–c usually appears in the first position of the third phrase of type 286.[15] By contrast, some melodic material is shared between types 252 and 277 (Figure 11), though never more than two subsequent phrases (out of four).[16] One of the most frequent short patterns, which is found in types 286, 252, and the full dataset, is the concluding 'cadence' b–a–g.[17] Entire melodies within our collection[18] are seldom transferable but may be found in other types, contexts, or folk song genres outside the scope of this collection. Further digitised collections are needed for more comprehensive exploration.

It is important to note that melodic transcriptions themselves are approximations; hence, examining such transcriptions and considering their absolute tone values—such as the collector, date, location, or performance conditions discussed below—may not be the most fruitful approach. A good example for this is the case of 'few-tone' melodies. Ethnomusicological research describes these as being of older historical origin than melodies with more than five tones. They were preserved in the song traditions recorded in the twentieth and twenty-first centuries, mainly in ritual songs and the musical traditions of peripheral and cross-border regions. As ballads were not typically used in ritual ceremonies, we will focus on the latter traditions. The musical characteristics of peripheral or cross-border regions, unlike those of central Slovenia, often reflect a blending with the musical characteristics of the folk music of neighbouring cultures. In some cases, these regions have (or had) limited exposure to broader musical influences due to cultural or geographic isolation. In our corpus, we identified a tone set for each song and

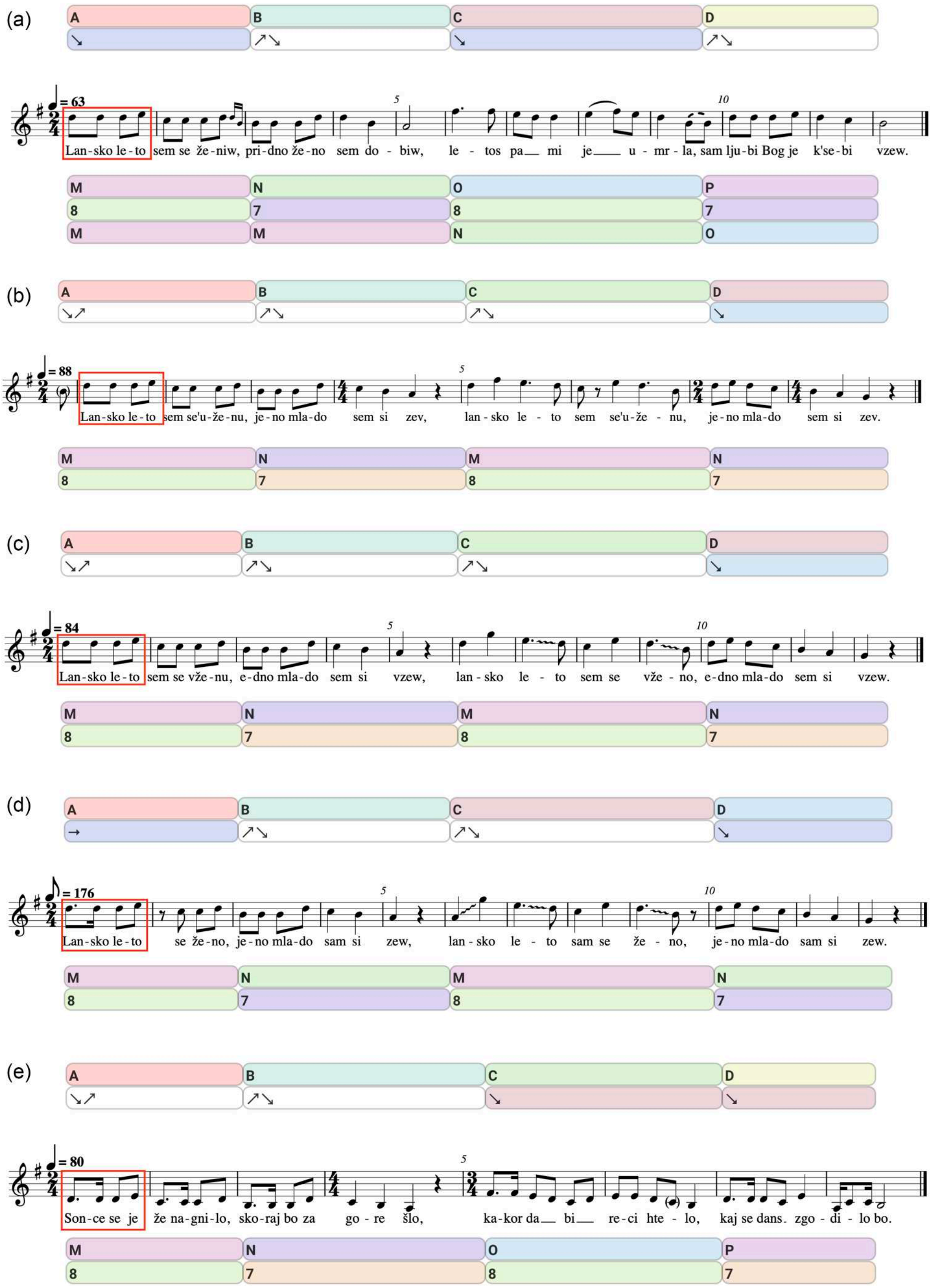

(a)
A B C D
↘ ↗↘ ↘ ↗↘
♩ = 63
5
10
Lan-sko le-to sem se že-niw, pri-dno že-no sem do - biw, le - tos pa__ mi je__ u - mr - la, sam lju-bi Bog je k'se-bi vzew.
M N O P
8 7 8 7
M M N O
(b)
A B C D
↘↗ ↗↘ ↗↘ ↘
♩ = 88
5
Lan-sko le-to sem se'u-že-nu, je-no mla-do sem si zev, lan-sko le - to sem se'u-že - nu, je-no mla-do sem si zev.
M N M N
8 7 8 7
(c)
A B C D
↘↗ ↗↘ ↗↘ ↘
♩ = 84
5
10
Lan-sko le-to sem se vže-nu, e-dno mla-do sem si vzew, lan-sko le - to sem se vže - no, e-dno mla-do sem si vzew.
M N M N
8 7 8 7
(d)
A B C D
→ ↗↘ ↗↘ ↘
♪ = 176
5
10
Lan-sko le-to se že-no, je-no mla-do sam si zew, lan-sko le - to sam se že - no, je-no mla-do sam si zew.
M N M N
8 7 8 7
(e)
A B C D
↘↗ ↗↘ ↘ ↘
♩ = 80
5
Son-ce se je že na-gni-lo, sko-raj bo za go - re šlo, ka-kor da__ bi__ re-ci hte - lo, kaj se dans. zgo - di - lo bo.
M N O P
8 7 8 7

Figure 11. The initial melodic phrase (pattern d-d-d-e), characteristic of type 252, is replicated with notable similarity (transposed an octave lower) in the identical position (initial) of the same phrase number (one) within the type variant 277.19. The figure features a selection of labelled type 252 variants (a) 113, (b) 121, (c) 125, and (d) 129, juxtaposed with e) a congruent variant 19 derived from type 277 (**a**) 252.113 (**b**) 252.121 (**c**) 252.125 (**d**) 277.19.

examined those containing three to five tones, i.e. tritonic to pentatonic melodies. The three-tone set was found in only one song, the four-tone set in two, and the five-tone set in 47. The following examples show how additional field material (metadata, sound recording, etc.) shed light on the processes that need to be considered when using the dataset for computational music analysis.

Firstly, a tritonic song may appear as result of a coincidence or an error in the data acquisition process. Alternatively, since an example of the song is sung by someone from Raba, a region in present-day Hungary inhabited by a Slovene minority, its 'few-tone' structure may reflect contact with neighbouring cultures. It could also reflect the preservation of older layers of musical tradition in relatively isolated communities. However, the audio recording and accompanying metadata reveal that the singer sang an additional verse that was not transcribed by the recorders. The second stanza is melodically different and more reminiscent of a rhythmic rendition of counting. This finding suggests that the song's apparent tritonic character is not inherent to the melody itself, but rather a result of incomplete transcription, underscoring the importance of consulting recordings and metadata when interpreting tone sets.

Secondly, certain pentatonic melodies (252.21/A3) exhibit similarities to heptatonic melodies (e.g. 252.24/A3). These melodies come from separate regions, one from Prekmurje and the other from Upper Carniola. It is plausible that the transcriptions of either melody may be flawed or oversimplified, or that they fail to account for singing errors, which often become apparent only upon hearing subsequent verses.[19] In another pentatonic example (257.6), the transcriber's note reveals that the singer performed only a fragment of the song, meaning the full melody was not captured.

Lastly, pentatonic melodies belonging to song type 256/A3, characterised by gradual movement and a narrow ambitus, come from the same melodic base, despite being recorded across diverse regions. Two cases are from Notranjska, and there are two from Styria, one from Lower Carniola, and one from Prekmurje. The ethnomusicologist Urša Šivic, in her commentary on examples of polyphony, notes that this melody

> began to appear more consistently in the second half of the twentieth century, across a broad ethnic territory (data for Carinthia and Littoral are not available). After 1956, as many as 33 out of 39 recorded melodies correspond to the aforementioned melody. (Šivic in Golež Kaučič et al. 2007: 345)

Conclusion

This article introduces our digitised dataset of Slovenian folk song family ballads and provides insight into factors shaping the current state of digitised music materials and their analysis. Drawing on Rice's five types of *theorising*, we introduce the corpus by addressing its *collection and conservation*, *organisation*, *structuring*, *explanation*, and *synthesis*.

To some extent, unique melodic, rhythmic and harmonic structures have evolved in folk music, independently of prevailing Western European musical conventions—perhaps due to social, cultural or geographical isolation. This is why understanding folk music and its cultural elements is both interesting and requires new approaches in music analysis, including computational approaches, which need to be tailored to the specificities of individual folk music cultures. Thus, grasping folk music and its cultural elements is crucial for music analysis, including that employing computational methods. Inspired by Cowdery (1984), our approach underscores the importance of evaluating new musical pieces in relation to others rather than in relation to abstract standards. This involves identifying overarching principles and flexible methods employed by practitioners, rather than imposing rigid external categories.

It is crucial to recognise the limitations inherent in computational principles when they are rooted in predetermined 'archetypes'. A comprehensive elucidation of the (meta)data is vital, alongside a critical evaluation of the ramifications of various types of labels applied to the data. This paper offers insights into the diverse factors influencing a particular dataset utilised for computer-assisted analysis, encompassing cultural, social, political, historical, and methodological dimensions. The computational analysis scrutinises metadata, descriptors, and melodic sequences. When confronted with extensive corpora, there is a tendency to examine overarching hypotheses. Our work straddled the realms of quantitative and qualitative inquiry, with findings frequently indicating that neither approach alone suffices (e.g. correlations between metadata and descriptors, analysis of 'few-tone' melodies). We underscore the significance of understanding one's dataset, emphasising the need to render digital data meaningful and accessible to broader audiences. Furthermore, we illustrate that quantitative findings warrant critical scrutiny, while qualitative insights often convey a richer narrative when derived from extensive datasets rather than focusing solely on individual cases.

Looking ahead, this work lays the foundation for several important avenues of development. A key objective was to establish a more suitable model for presenting digitised datasets, one extending beyond mere presentation of factual metadata and technical properties. This is particularly important in the case of materials such as these, for which relevant literature remains largely unavailable in languages other than Slovenian. The dataset and its accompanying documentation offer a valuable starting point for exploring the interplay between musical structure, lyrical content, and contextual information. Moreover, the detailed documentation enhances the potential for these materials to be employed in educational contexts, thereby contributing to the expansion of accessible corpora for teaching and learning. In addition, it allows the dataset to function as a testing ground for computational models, which can not only be evaluated but also meaningfully adapted to underrepresented repertoires such as this one. More broadly, the project encourages further interdisciplinary engagement, both locally and globally, and opens the door to a wide range of future applications in ethnomusicology, computational analysis, music preservation, music education, and other important objectives within the digital humanities.

Notes

1. See List's paper on 'manual' transcription agreement and adjustments of pitch and time duration in various folk songs (List 1974).
2. Originally, *Odbor za nabiranje slovenskih narodnih pesmi z napev*i = OSNP.
3. Transcribers needed thorough knowledge of regulations and had to complete their transcriptions rapidly in real time.
4. Contemporary researchers (see, for example, Golež Kaučič et al. 2007 (https://www.zotero.org/google-docs/?L88RO0)) acknowledge the drawbacks of such redaction, opting instead to highlight potential inaccuracies and hypothetical variants in footnotes of new editions, while ensuring the original records are accessible to the public.

5. For instance, guidelines encompassing interactions with subjects, audio capture methods, and embedding music within everyday life events were considered.
6. Different standards and protocols: https://www.loc.gov/librarians/standards, https://www.ifla.org/g/standards/current-ifla-standards/ [Accessed 2 April 2025].
7. ABC notation, https://abcnotation.com/ [Accessed first in winter 2023/4; at the time of submission, 20 March 2024, the link was not accessible].
8. '[…] a group of melodies showing basic interrelation by means of constant melodic correspondence, and presumably owing their mutual likeness to descent from a single air that has assumed multiple forms through processes of variation, imitation, and assimilation' (Bayard 1950: 33; see also Pendlebury 2020).
9. Dezrann (Slovenian folk song ballads), https://dezrann.net/explore/slovenian-folk-song-ballads [Accessed 2 April 2025].
10. For the sake of transparency, the title, first verse, and regional information have been translated into English by the authors of this publication. We acknowledge the possibility of bias introduced during the translation process. Scholars wishing to undertake lyrical or thematic analysis are advised to consult one of the custodians of the collection or an expert in the field of Slovenian folk songs or similar.
11. The file can be accessed through http://algomus.fr/data.
12. Josip Dravec straddles both the pre- and post-1950 eras, but the majority of his transcriptions were recorded after 1950, placing him firmly in the latter period. While some other transcribers also span both eras, the bulk of their work tends to be concentrated in one era, with only a few exceptions occurring in the other.
13. In the 'Collection and Conservation Practices' section, it is mentioned that before the 1940s, transcription methods were flawed due to the absence of recording equipment. These methods often conformed to Western art music standards, impacting melody (forced into major/minor scales), harmony (artificial harmonizations), and rhythm. As recording equipment improved, some 6/8 melodies were later found to have been sung in 5/8 or other (irregular) metres.
14. The files mentioned in this subsection can be accessed through http://algomus.fr/data and are a part of released dataset (Borsan et al. 2024).
15. The files mentioned in this subsection can be accessed through http://algomus.fr/data and are a part of released dataset (Borsan et al. 2024).
16. We tested the common starting patterns in 252 (g-g-g-d and d-d-d-e). Both aligned with positions in phrases of 277, matching first with variant 19 and second with variant 15. The latter also showed similarities with phrases one, three, five, and seven of 248.18.
17. The sequence appeared in 362 phrases. Among these, 313 instances were *not* found in the first two positions within a phrase, and 277 were *not* present in the first phrase.
18. Even similar melodic sequences are rarely fully duplicated or they are not duplicated at all. We found that no more than 25% of songs within the same type share identical melodic material.
19. Compare track 14 (Slovenske ljudske pesmi V 2007) and the notation for song 258.12.

Acknowledgements

We extend our heartfelt gratitude to Richard Groult and Baptiste Bacot for their invaluable contributions to this paper. Their rigorous review, insightful comments, and thoughtful suggestions have significantly enhanced the quality and clarity of our work. We are sincerely thankful for their time, expertise, and dedication to improving this manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by the Slovenian Research and Innovation Agency (ARIS) under the research programme P6-0111 *Folklore and Ethnological Research on Slovenian Folk Culture*, P5-0408 *Heritage on the Margins: New Perspectives on Heritage and Identity within and beyond National*, and J6-60108 *Identification of Slovenian Music after 1991*.

Notes on contributors

Vanessa Nina Borsan, PhD, is a Research Assistant at the Department of Musicology (Faculty of Arts, University of Ljubljana). She completed her bachelor's and master's degrees in Musicology at the University of Ljubljana and in Sound and Music Computing at Pompeu Fabra University, respectively. She recently completed her PhD at the University of Lille, working within the interdisciplinary Algomus research team. Her doctoral research explored intersections between computation and music, with a particular focus on developing computational methods for analysing musical structures and melodic patterns in Slovenian folk songs.

Mojca Kovačič, PhD, is a Research Associate at the Institute of Ethnomusicology of the Research Centre of the Slovenian Academy of Sciences and Arts (ZRC SAZU) in Ljubljana, Slovenia. Her ethnomusicological research focuses on musical and sonic phenomena, linking them to social, historical, political and other processes.

Mathieu Giraud is a CNRS Senior Researcher in the Centre de Recherche en Informatique, Signal et Automatique de Lille (CRIStAL), in Université de Lille, in France. He leads the Algomus computer music team, a Digital Humanities research group specialising in Musical Information Retrieval (MIR). The team focuses on high-level modelling, analysis, and co-creative generation of music.

Marjeta Pisk, PhD, is a Research Associate at the ZRC SAZU Institute of Ethnomusicology in Ljubljana, Slovenia. She intertwines folkloristic research of folk song texts and of other marginalised genres with research on heritage-making processes.

Matevž Pesek received a BSc degree in computer science and a PhD degree from the University of Ljubljana, in 2012 and 2018, respectively. He is currently an Associate Professor and a Researcher at the Faculty of Computer and Information Science, University of Ljubljana. He has been a member of the Laboratory of Computer Graphics and Multimedia, since 2009. His research interests include music information retrieval, music e-learning, biologically-inspired models, and deep architectures. He also researched compositional hierarchical modelling as alternative deep transparent architectures, and music multi-modal perception, including human-computer interaction in virtual reality, and visualisation for audio analysis and music generation.

Matija Marolt is a full Professor at the Faculty of Computer and Information Science at the University of Ljubljana where he is the Head of the Laboratory for Computer Graphics and Multimedia. His research interests include music/audio information retrieval, computer graphics, and visualisation. He focuses on problems such as melody and rhythm estimation, audio segmentation and organisation, and search and visualisation of music collections.

ORCID

Vanessa Nina Borsan http://orcid.org/0000-0001-8421-4409
Mojca Kovačič http://orcid.org/0000-0003-2496-6881
Mathieu Giraud http://orcid.org/0000-0003-2741-8047
Marjeta Pisk http://orcid.org/0000-0001-9350-7481
Matevž Pesek http://orcid.org/0000-0001-9101-0471
Matija Marolt http://orcid.org/0000-0002-0619-8789

References

Adams, Charles R. 1976. 'Melodic Contour Typology'. *Ethnomusicology* 20(2): 179–215.

Anagnostopoulou, Christina, Mathieu Giraud and Nick Poulakis. 2013. 'Melodic Contour Representations in the Analysis of Children's Songs'. Paper presented at the 3rd International Workshop on Folk Music Analysis. Meertens Institute and Utrecht University Department of Information and Computing Sciences, Netherlands. 6–7 June.

Antoniou, Grigoris and Frank van Harmelen. 2004. 'Web Ontology Language: OWL'. In *Handbook on Ontologies*, edited by Steffen Staab and Rudi Studer, 67–92. Berlin and Heidelberg: Springer.

Bayard, Samuel P. 1950. 'Prolegomena to a Study of the Principal Melodic Families of British-American Folk Song'. *The Journal of American Folklore* 63(247): 1–44.

Bendix, Regina. 1997. *In Search of Authenticity: The Formation of Folklore Studies*. Wisconsin: UW Press.

Bertin-Mahieux, Thierry, Daniel P. W. Ellis, Brian Whitman and Paul Lamere. 2011. 'The Million Song Dataset'. Paper presented at the 12th International Society for Music Information Retrieval Conference. University of Miami, FL, USA. 24–28 October.

Blaukopf, Kurt. 1993. *Glasba v družbenih spremembah: temeljne poteze sociologije glasbe*. Ljubljana: Škuc.

Boot, Peter, Anja Volk and W. Bas de Haas. 2016. 'Evaluating the Role of Repeated Patterns in Folk Song Classification and Compression'. *Journal of New Music Research* 45(3): 223–38.

Borsan, Vanessa Nina, Mathieu Giraud, Richard Groult and Thierry Lecroq. 2023. 'Adding Descriptors to Melodies Improves Pattern Matching: A Study on Slovenian Folk Songs'. Paper presented at the 24th International Society for Music Information Retrieval Conference. Politecnico Milano, Italy. 5–9 November.

Borsan, Vanessa Nina, Mojca Kovačič, Mathieu Giraud, Marjeta Pisk, Matevž Pesek and Matija Marolt. 2024. *The Digitised Dataset of Slovenian Folk Song Ballads (V1)* [Dataset]. *Recherche Data Gouv*. Retrieved from https://doi.org/10.57745/SINZFK

Bountouridis, Dimitrios, Daniel G. Brown, Frans Wiering and Remco C. Veltkamp. 2017. 'Melodic Similarity and Applications Using Biologically-Inspired Techniques'. *Applied Sciences* 7(12): 1242.

Bozkurt, Baris. 2015. 'Computational Analysis of Overall Melodic Progression for Turkish Makam Music'. In *Penser l'improvisation*, edited by Mondher Ayari, 289–98. Sampzon: Delatour France.

Caro Repetto, Rafael, Rong Gong, Nadine Kroher and Xavier Serra. 2015. 'Comparison of the Singing Style of Two Jingju Schools'. Paper presented at the 16th International Society for Music Information Retrieval Conference. Universidad de Málaga, Spain, 26–30 October.

Clayton, Martin, Preeti Rao, Nithya Nadig Shikarpur, Sujoy Roychowdhury and Jin Li. 2022. 'Raga Classification from Vocal Performances Using Multimodal Analysis'. Paper presented at the 23rd International Society for Music Information Retrieval Conference. Bengaluru, India. 4–8 December.

Conklin, Darrell. 2011. *Basque Songbook* [Dataset]. Retrieved from https://www.eusko-ikaskuntza.eus/en/documentary-collection/basque-songbook/

———. 2013. 'Antipattern Discovery in Folk Tunes'. *Journal of New Music Research* 42(2): 161–9.

———. 2021. 'Mining Contour Sequences for Significant Closed Patterns'. *Journal of Mathematics and Music* 15(2): 112–24.

Conklin, Darrell and Christina Anagnostopoulou. 2011. 'Comparative Pattern Analysis of Cretan Folk Songs'. *Journal of New Music Research* 40(2): 119–25.

Cornelissen, Bas, Willem Zuidema and John Ashley Burgoyne. 2021. 'Cosine Contours: A Multipurpose Representation for Melodies'. Paper presented at the 22nd International Society for Music Information Retrieval Conference. Online. 7–12 November.

Cowdery, James R. 1984. 'A Fresh Look at the Concept of Tune Family'. *Ethnomusicology* 28(3): 495–504.

dos Santos, Carolina L. and Carlos N. Silla. 2015. 'The Latin Music Mood Database'. *EURASIP Journal on Audio, Speech, and Music Processing* 2015(1): 23.

Eerola, Tuomas and Petri Toiviainen. 2004. *MIDI Toolbox: MATLAB Tools for Music Research*. Kopijyvä, Jyväskylä, Finland: University of Jyväskylä.

Etnofon. 2023. *Spletna zbirka Etnofon*. Glasbenonarodopisni inštitut ZRC SAZU. Retrieved from https://etnomuza.zrc-sazu.si/etnofon/ (Accessed 14 November 2025).

Floridi, Luciano. 2010. *Information: A Very Short Introduction*. Oxford: Oxford University Press.

Golež Kaučič, Marjetka. 2018. *Slovenska ljudska balada*. Ljubljana: Založba ZRC, ZRC SAZU.

Golež Kaučič, Marjetka, Marija Klobčar, Zmaga Kumer, Urša Šivic and Marko Terseglav. 2007. *Slovenske ljudske pesmi V: Pripovedni pesmi*. Ljubljana: Založba ZRC, ZRC SAZU.

Gómez-Cañón, Juan Sebastián, Nicolás Gutiérrez-Páez, Lorenzo Porcaro, Alastair Porter, Estefanía Cano, Perfecto Herrera-Boyer, Aggelos Gkiokas, et al. 2023. 'TROMPA-MER: An Open Dataset for Personalized Music Emotion Recognition'. *Journal of Intelligent Information Systems* 60(2): 549–70.

Hood, Mantle. 1982. *The Ethnomusicologist*. Kent: Kent State University Press.

Hoos, Holger H., Keith A. Hamel, Kai Renz and Jürgen Killian. 2001. 'Representing Score-Level Music Using the GUIDO Music-Notation Forma'. *Computing in Musicology* 12(1999–2000): 75–95.

Huang, Tongbo, Guangyu Xia, Yifei Ma, Roger Dannenberg and Christos Faloutsos. 2013. 'MidiFind: Fast and Effective Similarity Searching in Large MIDI Databases'. Paper presented at the 10th International Symposium on Computer Music Multidisciplinary Research. Marseille, France. 15–18 October.

Huron, David. 1996. 'The Melodic Arch in Western Folksongs'. *Computing in Musicology* 10(March): 3–23.

Killick, A. 2020. 'Global Notation as a Tool for Cross-Cultural and Comparative Music Analysis'. *Analytical Approaches to World Music* 8(2): 235–79.

Klobčar, Marija. 2014. '"Kako in kdaj so prišle žene do tega, da pojejo pesmi tudi o takih junakih, kakor sta Pegam in Lambergar?": Matija Murko in vprašanje nosilcev pripovednih pesmi'. *Glasnik Slovenskega etnološkega društva* 54(3): 21–9.

Kovačič, Mojca. 2015. *Glasbena podoba ljudske pesmi v rokopisnih, tiskanih in zvočnih virih v prvih desetletjih 20. stoletja*. Ljubljana: Založba Univerze v Ljubljani.

Kuhač, Franjo Ksaver. 1878–1881. *Južno-slovjenske narodne popievke*. 4 vols. Zagreb: Tiskara i litografija C. Albrechta.

Kumer, Zmaga. 1959. 'Slovenske ljudske pesmi z napevi: poročilo o glasbenem gradivu, nabranem 1906–1914 pod Štrekljevim vodstvom, zdaj v Glasbeno narodopisnem inštitutu v Ljubljani'. *Slovenski etnograf* 12: 203–10.

———. 1998. 'Pogled na dosedanje delo baladne komisije'. In *Ljudske balade med izročilom in sodobnostjo*, edited by Marjetka Golež, 31–2. Ljubljana: Založba ZRC, ZRC SAZU.

Kunej, Drago. 1999. 'Prva magnetofonska snemanja zvočnega gradiva Glasbenonarodopisnega inštituta'. *Traditiones* 28(2): 217–32.

———. 2005. '"We Have Plenty of Words Written Down; We Need Melodies!" The Purchase of the First Recording Device for Ethnomusicological Research in Slovenia'. *Traditiones* 34(1): 125–40.

———. 2022. 'Sound Recordings and Karel Štrekelj: The Initiator of a New Approach to Folk Song Research in Slovenia'. *Musicology* 33(December): 39–52.

Kunej, Drago and Rebeka Kunej. 2020. *Protokol za lastnike zvočnih zbirk*. Ljubljana: Založba ZRC, ZRC SAZU.

Lee, D. 2019. 'Hornbostel-Sachs Classification of Musical Instruments'. *Knowledge Organization* 47(1): 72–91.

Lidy, Thomas, Carlos N. Silla, Olmo Cornelis, Fabien Gouyon, Andreas Rauber, Celso A. A. Kaestner and Alessandro L. Koerich. 2010. 'On the Suitability of State-of-the-Art Music Information Retrieval Methods for Analyzing, Categorizing and Accessing Non-Western and Ethnic Music Collections'. *Signal Processing, Special Section: Ethnic Music Audio Documents: From the Preservation to the Fruition* 90(4): 1032–48.

List, George. 1974. 'The Reliability of Transcription'. *Ethnomusicology* 18(3): 353–77.

Lomax, Alan. 1978. *Cantometrics: An Approach to the Anthropology of Music*. California: UCL Extension Media Center.

Makris, Dimos, Ioannis Karydis and Spyros Sioutas. 2015. 'The Greek Music Dataset'. Paper presented at the 16th International Conference on Engineering Applications of Neural Networks. New York, NY, USA. 25 September.

McGuinness, Deborah L. and Frank van Harmelen. 2004. 'OWL Web Ontology Language Overview'. https://www.w3.org/TR/owl-features/. (Accessed 20 March 2024).

Mehr, Samuel A., Manvir Singh, Dean Knox, Daniel M. Ketter, Daniel Pickens-Jones, S. Atwood, Christopher Lucas, Nori Jacoby, Alena A. Egner, Erin J. Hopkins, Rhea M. Howard, Joshua K. Hartshorne, Mariela V. Jennings, Jan Simson, Constance M. Bainbridge, Steven Pinker, Timothy J. O'Donnell, Max M. Krasnow and Luke Glowacki. 2019. 'Universality and Diversity in Human Song'. *Science* 366(6468): eaax0868. https://doi.org/10.1126/science.aax0868

Merriam, Alan P. and Valerie Merriam. 1964. *The Anthropology of Music*. IL: Northwestern University Press.

Mills, Isabelle. 1974. 'The Heart of the Folk Song'. *MUSICultures* 2: 29–34.

Murko, Matija. 1929. 'Velika zbirka slovenskih narodnih pesmi z melodijami'. *Etnolog* 3: 5–54.

Neubarth, Kerstin and Darrell Conklin. 2016. 'Contrast Pattern Mining in Folk Music Analysis'. In *Computational Music Analysis*, edited by David Meredith, 393–424. Cham: Springer International Publishing.

Neubarth, Kerstin, Izaro Goienetxea, Colin G Johnson and Darrell Conklin. 2012. 'Association Mining of Folk Music Genres and Toponyms'. Paper presented at the 13th International Society for Music Information Retrieval Conference. Porto, Portugal. 8–12 October.

Nienhuys, Han-Wen and J. Nieuwenhuizen. 2003. 'Lilypond, a System for Automated Music Engraving'. Paper presented at the 14th Colloquium on Musical Informatics. Firenze, Italy. 8–10 May.

Nuttall, Thomas, Miguel García Casado, Víctor Núñez Tarifa, Rafael Caro Repetto and Xavier Serra. 2019. 'Contributing to New Musicological Theories with Computational Methods: The Case of Centonization in Arab-Andalusian Music'. Paper presented at the 20th International Society for Music Information Retrieval Conference. Delft, Netherlands. 4–8 November.

OSNP. 1907. 'Od "Odbora za nabiranje slovenskih narodnih pesmi"'. *Ljubljanski zvon* 27(3): 191–2.

Ossa, Sergio de la. 2019. *A Basic Guide to Folksong Analysis*. Budapest: Liszt Academy of Music.

Ozaki, Yuto, Adam Tierney, Peter Pfordresher, John Mcbride, Emmanouil Benetos, Polina Proutskova, Gakuto Chiba, Patrick E. Savage, et al. 2022. 'Globally, Songs Are Slower, Higher, and Use More Stable Pitches than Speech' [Stage 2 Registered Report]. https://osf.io/preprints/psyarxiv/jr9x7.

Ozaki, Yuto, Adam Tierney, Peter Q. Pfordresher, John M. McBride, Emmanouil Benetos, Polina Proutskova, Gakuto Chiba, Fang Liu, Nori Jacoby, Suzanne C. Purdy, Patrick E. Savage, et al. 2024. 'Globally, Songs and Instrumental Melodies Are Slower and Higher and Use More Stable Pitches than Speech: A Registered Report'. *Science Advances* 10(20): eadm9797. https://doi.org/10.1126/sciadv.adm9797

Papaioannou, Charilaos, Ioannis Valiantzas, Theodoros Giannakopoulos, Maximos Kaliakatsos-Papakostas and Alexandros Potamianos. 2022. 'A Dataset for Greek Traditional and Folk Music: Lyra' [Preprint]. https://arxiv.org/abs/2211.11479.

Pendlebury, Celia. 2020. 'Tune Families and Tune Histories: Melodic Resemblances in British and Irish Folk Tunes'. *Folk Music Journal* 11(5): 67–95.

Pesek, Matevž, Gregor Strle, Alenka Kavčič and Matija Marolt. 2017. 'The Moodo Dataset: Integrating User Context with Emotional and Color Perception of Music for Affective Music Information Retrieval'. *Journal of New Music Research* 46(3): 246–60.

Porter, Alastair and Xavier Serra. 2014. 'An Analysis and Storage System for Music Research Datasets'. Paper presented at the 1st International Workshop on Digital Libraries for Musicology. London, United Kingdom. 12 September.

Porter, Alastair, Mohamed Sordo and Xavier Serra. 2013. 'Dunya: A System for Browsing Audio Music Collections Exploiting Cultural Context'. Paper presented at the 14th International Society for Music Information Retrieval Conference. Curitiba, PR, Brazil. 4–8 November.

Proutskova, Polina, A. Volk, Peyman Heidarian and Gyorgy Fazekas. 2020. 'From Music Ontology Towards Ethno-Music-Onthology'. Paper presented at the 22nd International Society for Music Information Retrieval Conference. Online. 7–12 November.

Raimond, Yves, Samer Abdallah, Mark Sandler and Frederick Giasson. 2007. 'The Music Ontology'. Paper presented at the 8th International Society for Music Information Retrieval Conference. Vienna University of Technology, Austria. 23–27 September.

Ren, Iris Yuping. 2016. 'Closed Patterns in Folk Music and Other Genres'. In *FMA 2016*.

Ren, Iris Yuping, Anja Volk, Wouter Swierstra and Remco C. Veltkamp. 2018. 'Analysis by Classification: A Comparative Study of Annotated and Algorithmically Extracted Patterns in Symbolic Music Data'. Paper presented at the 19th International Society for Music Information Retrieval Conference. Paris, France. 23–27 September.

Renz, Kai. 2002. 'Algorithms and Data Structures for a Music Notation System Based on GUIDO Music Notation'. PhD diss., Technische Universität Darmstadt.

Rice, Timothy. 2017. *Modeling Ethnomusicology*. Oxford, New York: Oxford University Press.

Rizo, David and Alan Alexander Marsden. 2019. 'An MEI-Based Standard Encoding for Hierarchical Music Analyses'. *International Journal on Digital Libraries* 20(1): 93–105.

Rosenzweig, Sebastian, Frank Scherbaum, David Shugliashvili, Vlora Arifi-Müller and Meinard Müller. 2020. 'Erkomaishvili Dataset: A Curated Corpus of Traditional Georgian Vocal Music for Computational Musicology'. *Transactions of the International Society for Music Information Retrieval* 3(April): 31–41.

Rothstein, Joseph. 1992. *MIDI: A Comprehensive Introduction*. Madison, WI: A-R Editions, Inc.

Sapp, Craig Stuart. 2005. 'Online Database of Scores in the Humdrum File Format'. Paper presented at the 6th International Society for Music Information Retrieval Conference. London, United Kingdom. 11–15 September.

Savage, Patrick E. 2018. "Alan Lomax's Cantometrics Project: A Comprehensive Review". *Music & Science* 1(January): 1–19.

———. 2020. 'Measuring the Cultural Evolution of Music: Cross-Cultural and Cross-Genre Case Studies' [Preprint]. https://osf.io/mxrkw.

Savage, Patrick E. and Q. Atkinson. 2015. 'Automatic Tune Family Identification by Musical Sequence Alignment'. Paper presented at the 16th International Society for Music Information Retrieval Conference. Universidad de Málaga, Spain. 26–30 October.

Savage, Patrick E., Sam Passmore, Gakuto Chiba, Thomas E. Currie, Haruo Suzuki and Quentin D. Atkinson. 2022. 'Sequence Alignment of Folk Song Melodies Reveals Cross-Cultural Regularities of Musical Evolution'. *Current Biology* 32(6): 1395–402.e8.

Silla, Carlos Nascimento Jr., Alessandro L. Koerich and Celso A.A. Kaestner. 2008. 'The Latin Music Database'. Paper presented at the 9th International Society for Music Information Retrieval Conference. Drexel University, PA, USA. 14–18 September.

Singh, Yeshwant, Lilapati Waikhom, Vivek Meena and Anupam Biswas. 2022. *Indian Folk Music Dataset* [Dataset]. *Zenodo*. Retrieved from https://doi.org/10.5281/zenodo.6584021

Slovenske ljudske pesmi V, Družinske pripovedne pesmi. 2007. *From the Archives of the Institute of Ethnomusicology*. Ljubljana: Založba ZRC, ZRC SAZU.

Srinivasamurthy, Ajay, Sankalp Gulati, Rafael Caro Repetto and Xavier Serra. 2021. 'Saraga: Open Datasets for Research on Indian Art Music'. *Empirical Musicology Review* 16(1): 85–98.

Stefanija, Leon, Vanessa Nina Borsan, Matevž Pesek, Matija Marolt, Drago Kunej and Zoran Krstulović. 2022. 'Zgodovina in izzivi digitalne etno/muzikologije v Sloveniji'. *Musicological Annual* 58(2): 15–49.

Štrekelj, Karel. 1906. *Navodila in vprašanja za zbiranje in zapisovanje narodnih pesmi, narodne godbe, narodnih plesov in šeg, ki se nanašajo na to*. Ljubljana: Zadružna tiskarnica.

Strle, Gregor and Matija Marolt. 2012. 'The EthnoMuse Digital Library: Conceptual Representation and Annotation of Ethnomusicological Materials'. *International Journal on Digital Libraries* 12(2–3): 105–19.

Temperley, David. 2000. 'Meter and Grouping in African Music: A View from Music Theory'. *Ethnomusicology* 44(1): 65–96.

Terseglav, Marko. 1987. *Ljudsko pesništvo*. Ljubljana: Državna založba Slovenije.

Tian, Mi, György Fazekas, Dawn Black and Mark Sandler. 2013. 'Towards the Representation of Chinese Traditional Music: A State of the Art Review of Music Metadata Standards'. Paper presented at the International Conference on Dublin Core and Metadata Applications. Lisbon, Portugal. 2–6 September.

Urkizu, Izaro Goienetxea, Iñaki Arrieta Urtizberea, J. Bagüés, Arantza Cuesta, Pello Leiñena and D. Conklin. 2012. *Ontologies for Representation of Folk Song Metadata*. San Sebastian: University of the Basque Country. https://addi.ehu.es/bitstream/handle/10810/8053/tr12-1.pdf. (Accessed 20 March 2024).

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

Van Kranenburg, Peter and Folgert Karsdorp. 2014. 'Cadence Detection in Western Traditional Stanzaic Songs Using Melodic and Textual Features,' Paper presented at the 8th International Society for Music Information Retrieval Conference. Taipei, Taiwan. 27–31 October.

Van Kranenburg, Peter, Anja Volk and Frans Wiering. 2013. 'A Comparison between Global and Local Features for Computational Classification of Folk Song Melodies'. *Journal of New Music Research* 42(1): 1–18.

Van Kranenburg, Peter, Anja Volk, Frans Wiering and Bente Maegaard. 2011. 'On Operationalizing the Musicological Concept of Tune Family for Computational Modeling'. In *Supporting Digital Humanities: Answering the Unaskable*, edited by Bente Maegaard.

von Hornbostel, Erich M. and Curt Sachs. 1914. 'Systematik Der Musikinstrumente. Ein Versuch'. *Zeitschrift Für Ethnologie* 46(4/5): 553–90.

Vodušek, Valens. 2003. *Etnomuzikološki Članki in Razprave*. Edited by Marko Terseglav and Robert Vrčon. Ljubljana: Založba ZRC.

Volk, Anja, W. Bas de Haas and Peter Van Kranenburg. 2012. 'Towards Modelling Variation in Music as Foundation for Similarity'. Paper presented at the joint conference of MPC 2012 and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music. Thessaloniki, Greece. 23–28 September.

Volk, Anja, Jörg Garbers, Peter Van Kranenburg, Frans Wiering, Remco C Veltkamp and Louis P Grijp. 2007. 'Applying Rhythmic Similarity Based on Inner Metric Analysis to Folksong Research'. Paper presented at the 8th International Conference on Music Information Retrieval. Vienna, Austria. 23–30 September 2007.

Volk, Anja and Peter van Kranenburg. 2012. 'Melodic Similarity among Folk Songs: An Annotation Study on Similarity-Based Categorization in Music'. *Musicae Scientiae* 16(3): 317–39.

Weigl, David M., Tim Crawford, Aggelos Gkiokas, Werner Goebl, Emilia Gómez, Nicolás F. Gutiérrez, Cynthia C. S. Liem and Patricia Santos. 2021. 'FAIR Interconnection and Enrichment of Public-Domain Music Resources on the Web'. *Empirical Musicology Review* 16(1): 16–33.

Weigl, David M., Werner Goebl, Tim Crawford, Aggelos Gkiokas, Nicolas F. Gutierrez, Alastair Porter, Patricia Santos, et al. 2019. 'Interweaving and Enriching Digital Music Collections for Scholarship, Performance, and Enjoyment'. Paper presented at the 6th International Conference on Digital Libraries for Musicology. National Library of The Netherlands. 9th November. 84–8.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'. *Scientific Data* 3(1): 1–9.

Wood, Anna L. C., Kathryn R. Kirby, Carol R. Ember, Stella Silbert, Sam Passmore, Hideo Daikoku, et al. 2022. 'The Global Jukebox: A Public Database of Performing Arts and Culture'. *PLoS One* 17(11): e0275469.

Zembylas, Tasos. 2012. *Kurt Blaukopf on Music Sociology – An Anthology*. Frankfurt am Main: Peter Lang.