https://doi.org/10.7592/FEJF2024.94.finnic_corpus

BRIDGING NORTHERN AND SOUTHERN TRADITIONS IN THE FINNIC CORPUS OF ORAL POETRY

Kati Kallio
Academy Research Fellow
Finnish Literature Society
Docent in Folklore Studies
University of Helsinki, Finland
kati.kallio@helsinki.fi

Mari Sarv
Leading Research Fellow
Estonian Folklore Archives
Estonian Literary Museum, Estonia
mari@haldjas.folklore.ee

Maciej Janicki
Postdoctoral Researcher
Department of Digital Humanities
University of Helsinki, Finland
maciej.janicki@helsinki.fi

Eetu Mäkelä
Professor of Digital Humanities
University of Helsinki, Finland
eetu.makela@helsinki.fi

Abstract: Historical Finnic oral poetry – called *runolaulu*, *regilaul*, or Kalevalaic poetry – makes a versatile corpus across several related languages, representing numerous genres from epic and charms to lyric, ritual poetry, lullabies, and so forth. Despite the first comparative efforts and collaborations already at the end of the nineteenth century, the local traditions in Northern and Southern Finnic languages have mostly been analysed within national research traditions.

Data-driven approaches have a potential to reveal new perspectives to this multilingual tradition, especially to the less studied parts, and the overall characteristics of it. Yet, due to the multilevel linguistic, poetic, and cultural variation of the data, the use of computational methods is complicated and, typically, necessitates interlaying quantitative analyses with close reading and source-critical approaches.

In this paper, we introduce some results at the intersection of Northern and Southern Finnic song text corpora discovered with the help of similarity detection analyses. Our approach consolidates the idea of the complex interplay of divergence and commonality of regional runosong traditions. While often having particular features not found in other parts of the Finnic area, the regional traditions are also connected to one another by similar formulas, motifs, poem types and themes, and, at the same time, distinct in their variations, uses and interpretations of these. We hope that our tools also help others in examining these further.

Keywords: oral poetry, runosongs, Finnic languages, variation, verse similarity

Most Finnic languages share an oral tradition in a poetic system that is featured by irregular alliteration and parallelism, and a specific metre with a trochaic core. This tradition is currently called by several names, such as *regilaul* in Estonia and *runolaulu* or “Kalevalaic poetry” in Finland and Karelia.[1] Here, we use the term *runosong* for the tradition with all its branches and variants. With local variations, a similar poetic system has been in use in most of the Finnic languages. It has lived side by side with other metrical systems, such as free verse laments, children’s songs with shorter lines, free verse and prose charms and, during recent centuries, stanzaic rhymed songs of various kinds.

The notion that the oral traditions in Northern and Southern Finnic languages share not only a similar poetic system but also some content is an old one.[2] The interconnections derive from the common origins of the Finnic languages and poetic tradition, and from complex processes of diversification and interactions ever since (see, e.g., Grünthal 2020; Kallio 2015; Lang 2016; Korhonen 1994; Frog 2019). However, wide systematic up-to-date surveys of connections between the poetic cultures in the Finnic languages are lacking for several reasons. The archival data is wide. The local Finnic traditions have mostly been curated into and analysed as three separate national collections in Estonia, Russian Karelia, and Finland, typically within parallel research traditions. In practice, no one knows the whole data well enough to make encompassing comparisons: efforts have mostly been made at the level of individual song types and motifs, especially in the early twentieth century research (e.g., Krohn 1903–1910; 1931).

Currently, large amounts of poems from Estonian and Finnish archival collections are available in digital text form.[3] This makes it much easier to browse the texts (see Kallio & Mäkelä 2019; see also Lintrop 2024; Seppä 2021; Sykäri 2020) or to apply and develop computational analysis methods (e.g., Sarv & Järv 2023; Sarv & Kallio & Janicki 2024). Indeed, the computational methods for the analysis of folklore are quickly developing (see, e.g., Abello et al. 2023; Eklund & Hagedorn & Darányi 2023; Tangherlini et al. 2020).

The data itself, however, poses challenges for digital and quantitative approaches: the long recording and curating history of a complex and varying set of poetic cultures did not produce a user-friendly or balanced corpus (Kallio et al. 2023; see also Ilyefalvi 2020). Linguistic, poetic and orthographic variation of the corpus is extensive, the available dialectal dictionaries do not cover the whole corpus, there are no parsers or language models that fit the data, and, despite the considerable size, the data appears to be too small and heterogenous for machine learning or creating language models. This means that, for example, some truly inspiring approaches to compare or recognize poems and narrative traditions across relative languages (e.g., Jänicke & Wrisley 2017; Meder & Himstedt-Vaid & Meyer 2023; Meinecke & Wrisley & Jänicke 2021) are not, for now at least, feasible for runosong data. Although the Estonian and Finnish corpora are built up following similar metadata structures, the type indices used in Finnish and Estonian databases have different structures and characteristics. Thus, in this current state, it is truly hard to track down the common themes and motifs among the vast number of texts and song types in Southern and Northern Finnic corpora. By Southern Finnic languages we mean North and South Estonian (including Seto) and Votic, while Northern denotes Ingrian (Izhorian), Karelian and Finnish. With the exception of Votic, the runosongs in Southern Finnic languages are mostly included in the Estonian ERAB corpus and the runosongs in Northern Finnic languages in the Finnish SKVR and JR corpora.

In the interdisciplinary FILTER project,[4] we have explored computational means suitable for quantitative and qualitative analysis of the runosong data, and for interleaving these. We combined different runosong datasets into a joint SQL[5] database, set the data into a text and metadata search interface Octavo, and created a base map that fits the historical parish information of the data. We have been concentrating on methods that are suitable for our non-standard language data, where lines with the same content words exhibit significant dialectal, morphological, poetic, and orthographic variation, especially on methods for similarity recognition of poetic lines and for alignment and comparison of passages and texts (Janicki & Kallio & Sarv 2023; Janicki 2023). These similarity recognition results are offered for close reading in similarity recognition interface Runoregi, which enables qualitative exploration of similar lines, passages, and texts side by side, as clusters, in dendrogram or on the map (Janicki et al. 2024).[6] The bigram-based similarity recognition method works surprisingly well in identifying similarities in regional corpora or songs from nearby linguistic areas, and for recognising oral sources for literary works and verse-level effects of runosong publications on local oral cultures. This is due to the formulaic and repetitive variations of the poetic idiom. Yet, although it is known that there are similar poem types, motifs, and formulas also in the more distant areas, these are much more difficult to recognize computationally, due to greater content, motif and formula variation caused by partly linguistic, partly cultural distances.

In this article, we introduce some explorations on the similarities between texts in the Estonian language ERAB corpus and Karelian-Ingrian-Finnish language SKVR corpus, discovered with the help of different similarity detection analyses. Although the similarities between Northern and Southern Finnic traditions typically do not take place at the level of formulaic similarity of longer motifs or entire texts, the verse similarity detection method can bring together smaller elements such as similar verse types, which can then help in bringing together longer elements and contents that are – in terms of language, formulas, and verse types – too distant to be identified by using bigram similarity. The analysis catches verse types, which are shared in Southern and Northern Finnic oral poetry and contain similar word stems. We also compare different methods with one another and with the existing ERAB and SKVR poem type indices.

BASIC PROPERTIES OF THE THREE RUNOSONG CORPORA

The joint database of Finnic runosongs, compiled within the FILTER project, contains three large text corpora: the database of Estonian runosongs (ERAB, Eesti regilaulude andmebaas), the Finnish SKVR corpus (digitized version of the book series *Suomen Kansan Vanhat Runot* 'The old poems of the Finnish [Finnic] people') and the Finnish JR (Julkaisemattomat runot) corpus of unpublished poems.[7] Curating, typologizing, publishing, and digitizing the data has been a long effort of numerous people at the Estonian and Finnish folklore archives.

1) ERAB (the Database of Estonian Runosongs) of the Estonian Folklore Archives currently consists of 108,968 texts and is a work in progress: the database is constantly supplemented with texts from archival collections. The data is for the most part extracted from the OCR[8] of earlier machine-typed copies transcribed from manuscripts. The texts are checked with original manuscripts, and an orthographically normalized version is added. In addition to runosongs, the database contains some texts of other genres – for example, older-type dance or game songs, children's songs that may have some connection to runosong tradition, and the information on performance and performers of runosongs, making up approximately one quarter of the items in the database. The type

names retrieved from the machine-typed copies were assigned throughout several decades and are not in a totally coherent system. Hence, the pressing need to revise typology, which is in progress. Currently, the revised index contains 2,514 type names, and the original unsystematized one – with all the name versions – 14,818. Of these, 828 type names in the new index and 10,181 in the old one are only given to one poem text.[9] In this article, we use the revised type index, which does not cover the whole corpus yet (see Fig. 1). The corpus mainly contains material in Estonian languages and dialects.

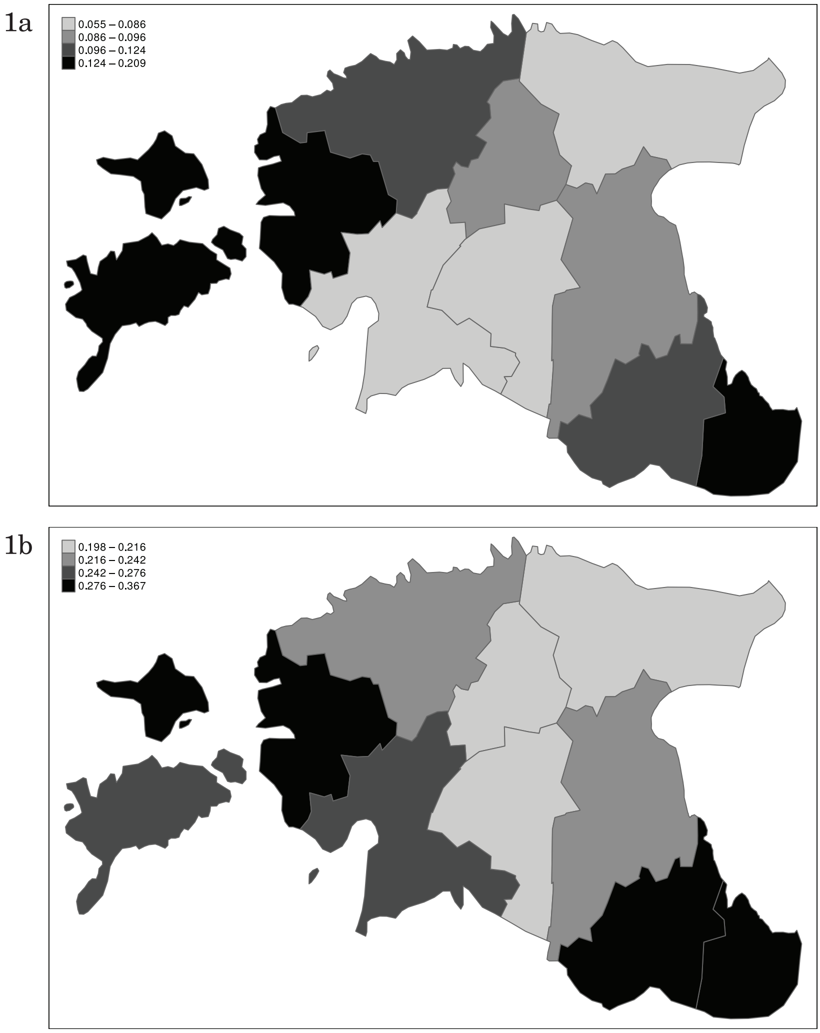


Figure 1. *Share of texts a) with missing type names, and b) classified to other genres than runosong in the Estonian ERAB corpus.*

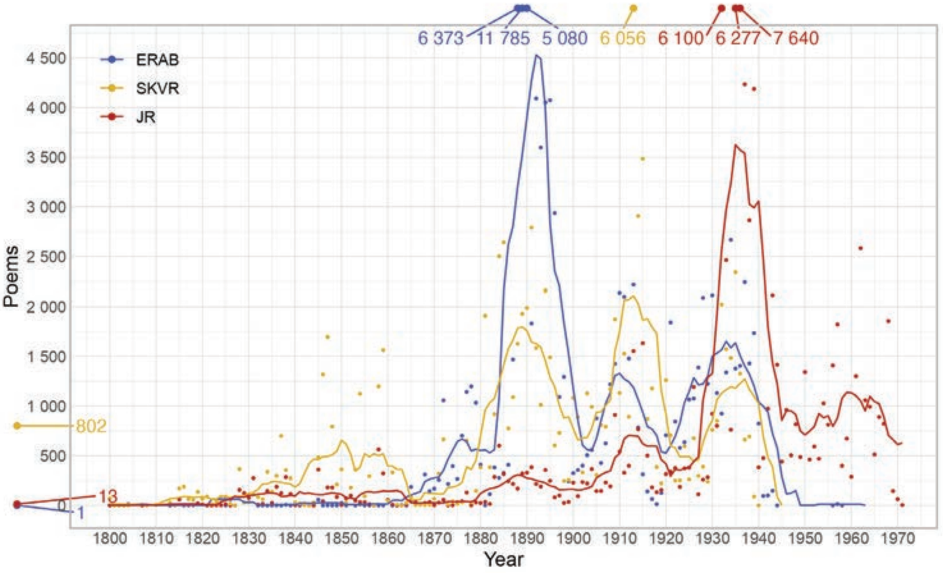


Figure 2. *The recording timelines of the three runosong corpora during the 19th–20th centuries. In addition, ca. 800 texts are from the 17th–18th centuries. The spots represent yearly counts, the lines five-year averages counted based on these. The two spots above the plot with numbers represent two years with exceptionally big collections in the ERAB and the JR, and also affect the five-year averages.*[10]

2) SKVR is the corpus of Karelian, Ingrian, Votic, and Finnish runosongs, originally in archival manuscripts, edited and published as a book series (1908–1948, 1997) and then digitized in the early 2000s. It consists of 89,247 texts and has a type index created using the old volume-specific indices, partly re-analysing the corpus, compiled from the 1980s onwards at the Folklore Archives of the Finnish Literature Society. The index contains 7,573 type names, of which 2,827 are indexed to one poem text only.

3) The JR corpus (Julkaisemattomat runot ‘Unpublished Poems’) contains 85,228 texts mostly from Finland, Karelia, and Ingria, also 1,142 texts from the Veps area and 12 from Setomaa. It is a selection of poems that were not published in or were recorded after the publication of the SKVR book series, later copied into a card file and digitized into a corpus of Unpublished Poems. The JR contains a significant number of other than runosong texts: children’s songs, rhymed songs, songs of literary origin, and even some Sami joiks and Karelian laments. While all corpora in our joint database contain some data other than runosongs, the number of other genres is highest in the JR. The corpus also contains copies – especially the oldest collections often circulated

among scholars, each making their own copies and editions. The metadata of the JR is not verified, and it contains no type index. Due to the very heterogeneous and non-verified nature of the JR corpus, and the lack of type index, we do not use this corpus in the explorations of the present article.

Having these three corpora in one database allows basic quantitative analysis of the data. All in all, our database combining these three corpora consists of 283,568 texts, mostly recorded during the late nineteenth and early twentieth centuries (Fig. 2). These contain 14,204,631 words and 1,084,389 unique word forms. The timeline of the accumulation of the collections shows that the three peaks have occurred around the same time in Estonian and Finnish collections, which would warrant a more detailed analysis.

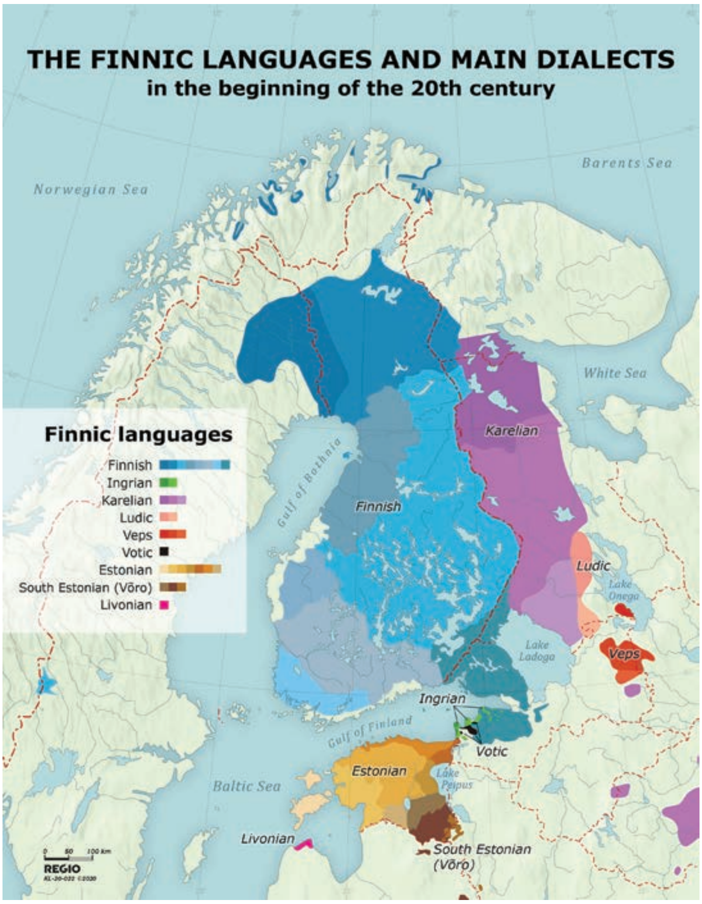
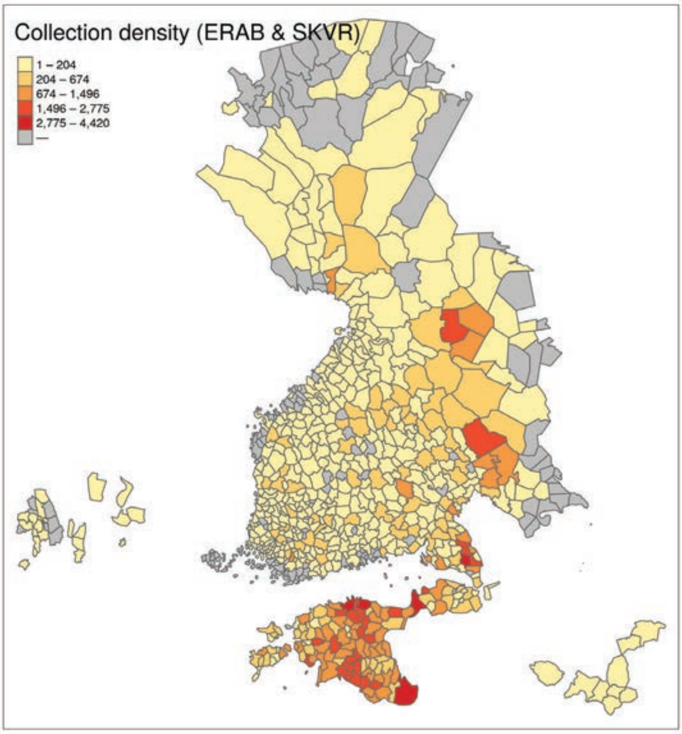


Figure 3. *Map of the Finnic languages and dialects at the beginning of the 20th century (Grünthal 2020: 6).*

The Estonian runosong corpus consists mainly of songs in North- and South-Estonian dialects, and the Finnish one of songs in Karelian, Ingrian, Votic, Ludic, and Finnish languages and their dialects. As languages are not coded in the data – and the borders of languages are often ambiguous – the number of texts in different languages is not reachable but rather constitutes a complex research question for future work. In terms of language history, Southern Finnic languages – South and North Estonian and Votic – are near to one other, and the Northern Finnic languages – Karelian, Ingrian, Ludic, and Finnish – form another group of close-by varieties, but the languages are also tied by their close geographical proximities and long interaction (e.g., Ingrian, Ingrian-Finnish and Votic in Ingria). In this article, due to our current metadata, we are not treating Votic, Ingrian, and Ingrian-Finnish songs as separate categories.

The map of the Finnic languages and their dialects at the beginning of the twentieth century (Fig. 3; Grünthal 2020) is a good reference to the nineteenth- and twentieth-century runosong collections, although the earlier immigrations, assimilations, and substrates affecting the local poetic cultures are not shown, and the complexity of multilingual areas especially in Ingria and Ladoga Karelia is difficult to visualize.



4a

4b

4c

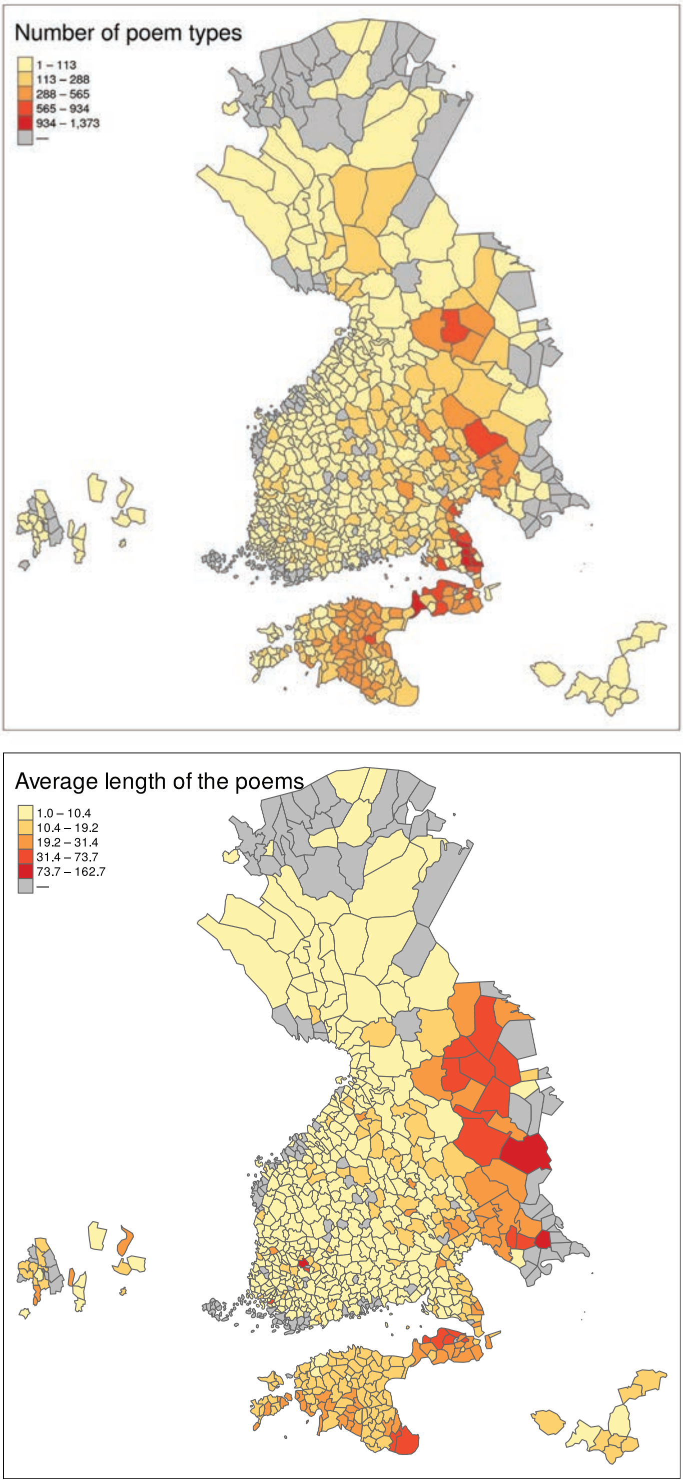


Figure 4. *a) Collection density, i.e., the number of poems collected from each parish; b) the number of different poem types by parishes in the ERAB and SKVR corpora; and c) average length of the poems. A minor part of the texts lacks the parish level indications or are only given the county level information.*

In the recorded collections, the number of poems, the average length of poems, and the number of indexed poem types in the material roughly follow similar patterns. More poems have been recorded from northern Estonia, Setomaa, and particular Ingrian and Karelian areas than elsewhere. On average, the poems are longest in Setomaa, coastal Ingria, and Karelian language areas. The amount of poem types in our indices is largest in Ingria, Karelian Isthmus, and Karelian language areas. These areas have versatile lyric song traditions, which in the SKVR corpus have been indexed at the level of small motifs, which affects the maps. (Fig. 4a–c.) In addition to the recording history and local poetic cultures, also the share of indexed material and other genres in the Estonian corpus (Fig. 1a–b) affects the analyses or poem types and genres.

At the time of recording, the local oral cultures were different, affected by various local and ethnic histories, linguistic and cultural exchanges, local livelihoods, political and religious developments, and processes of literacy and modernization (see Siikala 1994 and Sarmela 1994 for Northern Finnic areas). Further, the interest of the scholars and others in documenting the traditions, especially during the nineteenth century, focused on particular genres, poem types, and poetics (particularly on archaic mythological epics in regular poetic meter), and following these interests, on regions where the tradition corresponded best to the scholarly or elite ideologies, such as Karelian language area or Setomaa (see, e.g., Kalkun 2015; Sarv 2012; Tarkka 2013). The interests also changed in time.

Early nineteenth-century recorders mostly went to eastern Finland and Karelia, searching especially for long and archaic mythological epics and charms, which were available there. During the second part of the century, the overall focus widened to the whole Finnic area and a much wider set of genres. During the early twentieth century, the aim was to cover better also the areas in western Finland, which resulted in collecting large amounts of songs for children and very short charms, as these were at the time the most well-known runosong genres in the area. In Estonia, the collection process produced more even data, although the northern Estonian coast, Mulgimaa in the south, and Setomaa in the south-east are known as versatile tradition areas in general and have also more runosong texts collected.

These kinds of reasons explain the data distributions especially in the Northern Finnic areas, also in terms of recorded poetic genres (see below). The maps and plots of the work of individual recorders, and the distributions of materials collected from individual parishes at different times enable a much more detailed view into the collecting history and the local singing cultures than we can treat here or than has yet been written into articles (Kallio et al. 2023; Mäkelä & Kallio & Janicki 2024).

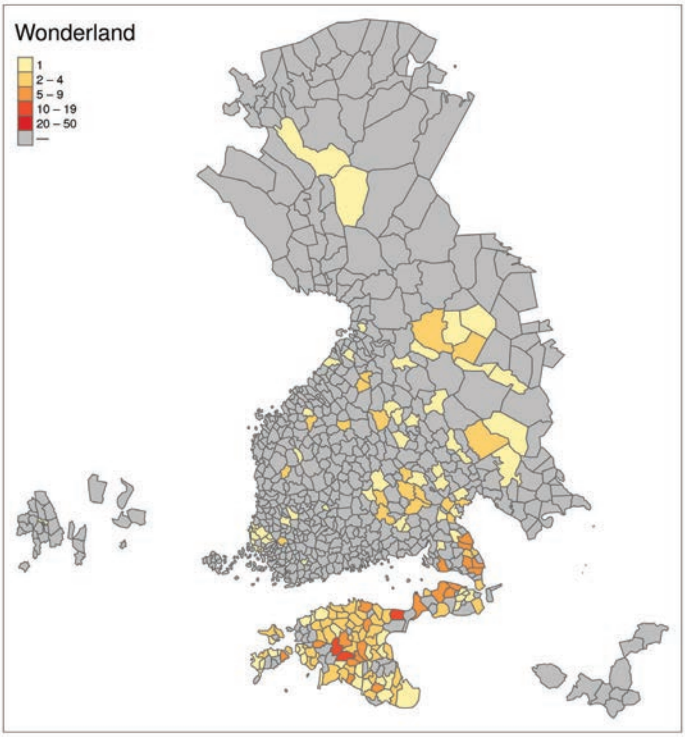
SHARED SONG TYPES IN ESTONIAN AND FINNISH INDICES

The comparison and typologization efforts of Finnish (Karelian, Ingrian, and Finnish) and Estonian materials have a long and intertwined history. The late nineteenth- and early twentieth-century perspectives to folklore emphasized the need for international comparisons, and folklorists in Estonia and Finland – citizens of the same Russian Empire – also had personal ties. At the end of the nineteenth century the geographical-historical method developed in folkloristics, with the main aim to discover the original form and place of folklore types by comparing text variants. Finnish researcher Julius Krohn was the main developer of the method, and his son Kaarle formulated the methodological guidelines for its application, runosongs being an important focus of interest for both. The method had a considerable impact on collecting folklore, also in Estonia, pointing to the necessity of meticulously writing down also quite similar versions of poems and tales, and to paying attention to the details of expression. Although many basic premises of the theory have since been questioned, it has produced a wide range of comparative literature on the variations of Finnic poem types (see, e.g., Frog 2021; Hautala 1954).

The geographical-historical method also called for international comparisons of runosongs and led to cross-Finnic research initiatives (Krohn 1903–1910, 1931). The Soviet period brought the collaboration to a halt, but already in the 1980s, Ülo Tedre and Matti Kuusi discussed the subject and even did some preparatory work for a wider project (Kuusi & Tedre 1979; SKSÄ 2023:3; SKSÄ 2023:30; Virtanen 1987: 32). Leea Virtanen (1987) summarized the research done in the second half of the twentieth century (mostly outside of the Soviet Union though) about the relationships of different Finnish, Karelian, Ingrian, and Estonian folklore genres. In the early 2000s, both Estonian and Finnish collections were digitized by Arvo Krikmann and his team within a collaboration project in Tartu, with an idea of enabling also further comparative work (Harvilahti 2013, 2019; Saarinen 2006; Sarv & Oras 2020; SKSÄ 2023:3; SKSÄ 2023:30).

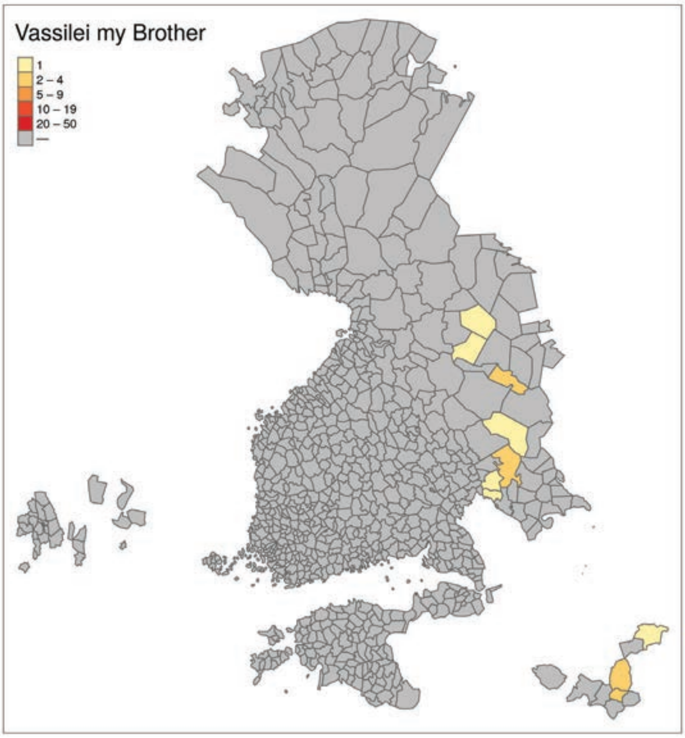
The recording efforts, indexing of the archival materials, publication projects, and the research of Finnic runosongs were tightly interwoven, and this work still echoes in the typologies currently in use. In the ERAB, the current harmonized index mostly follows the typology by Ülo Tedre and others in the anthology of Estonian runosongs *Eesti Rahvalaulud: Antoloogia* (1969–1974), with mainly thematic categorization principle. In the SKVR, criteria are distinct for different genres: lyrical poems are analysed by quite small motifs, ritual poems by ritual context and motifs, and charms by the function, if the knowledge is available. The epic index derives closely from the older indices, and is analysed as wider plots and themes. Thus, besides the complex oral variation, the internal variation of the typology poses challenges to the computational approaches. At the same time, the work of earlier scholars of comparing and organizing the data into publications and card files with type indices is the groundwork enabling further analysis.

After bringing the Finnish and Estonian data into a joint database, our first experiments in finding similar poem types and motifs between the SKVR and ERAB were simple and half-manual, just combining a list of the most frequent poem types and starting to go through a small part of it with the experts of both corpora, looking for potentially similar types, such as “Big Oak”, “Four Maidens”, or “Wonderland”. We also used earlier research and checked potential candidate pairs with text and metadata queries (with both direct SQL queries and Octavo interface; see Kallio & Mäkelä 2019). This kind of work has the potential to also identify similarities that do not occur at the level of verses or formulas but are vaguer and more thematic. Individual poem types may have very different reaches and warrant future research (see Fig. 5.).

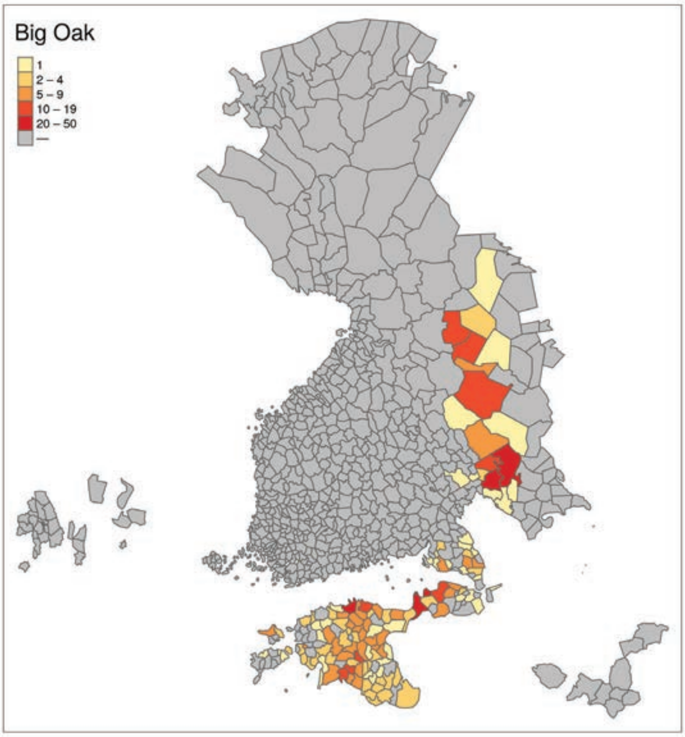


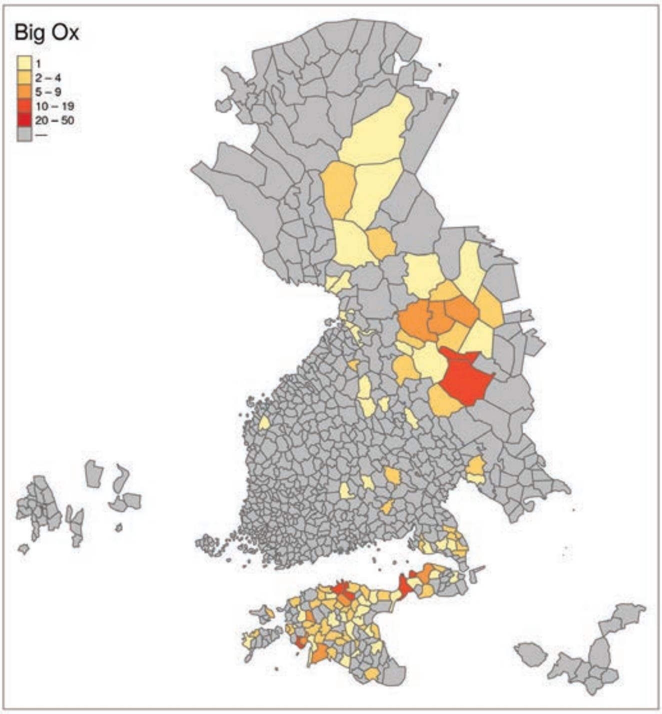
5a

5b

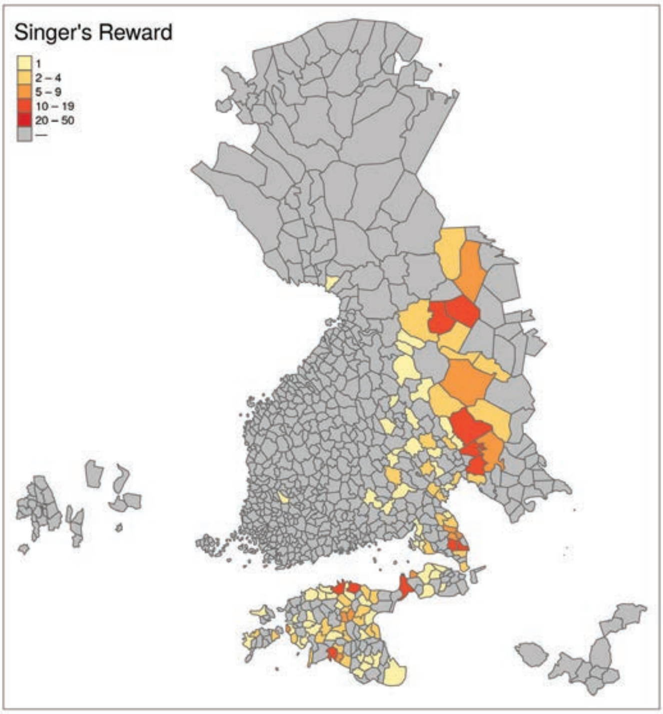


5c





5d



5e

5f

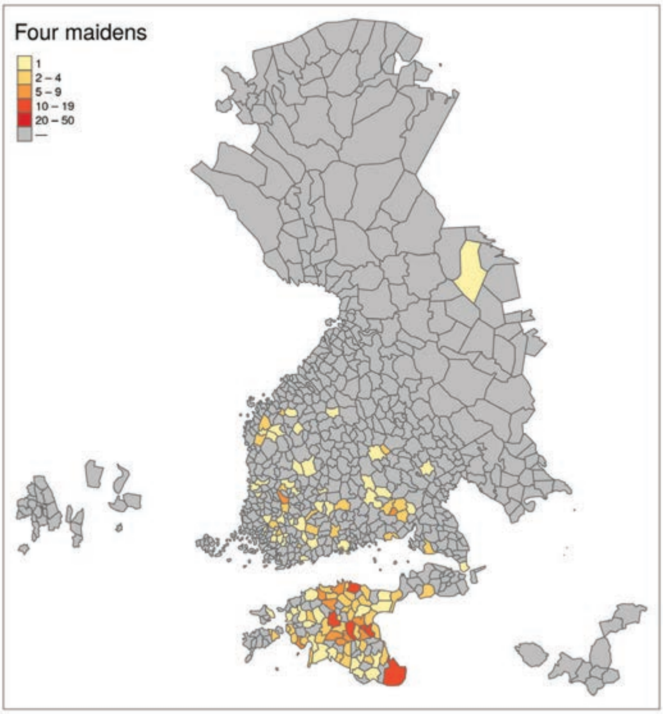
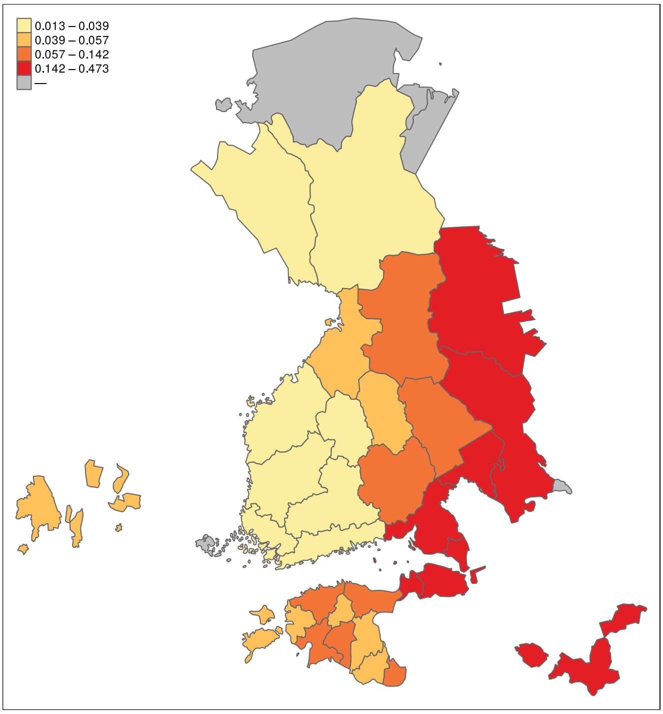
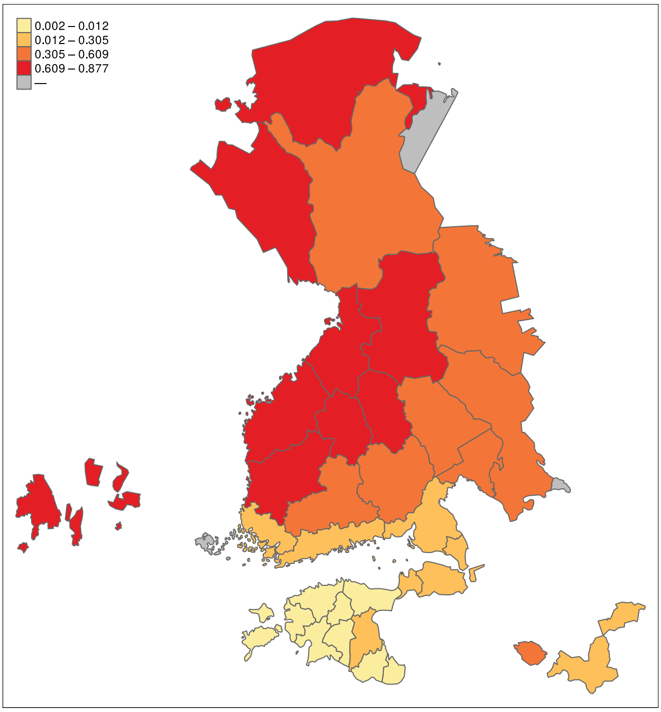


Figure 5. *The texts with parish-level place information indexed in the ERAB and SKVR with the title a) "Wonderland"; b) "Vassilei My Brother"; c) "Big Oak"; d) "Big Ox"; e) "Singer's Reward"; f) "Four Maidens".*

The poem type distributions may cover parishes in all the main language areas (e.g., Fig. 5a) or be limited to one language area only (e.g., Fig. 5b) or to some parishes in the case of more local poem types. For the types shared across the Northern-Southern Finnic linguistic border, the spreads most often include Estonia, Ingria, and Karelia (e.g., Fig. 5c): this probably relates to different processes linked to Christianization, Lutheran Church, literarization and modernization in Finland, and the wide popularity of Lutheran hymns and popular rhymed songs there in the nineteenth century. Yet, in many cases (e.g., Fig. 5d, 5e) the densest collections or longest versions come from Estonian, Ingrian, and Karelian language areas, but cases from the area of contemporary Finland show the type has been in use also there. Some poem types (e.g., Fig. 5f) are recorded mostly from western Finland and Estonia, or in some other regional combinations.


0.013 – 0.039
0.039 – 0.057
0.057 – 0.142
0.142 – 0.473
—

6a


0.002 – 0.012
0.012 – 0.305
0.305 – 0.609
0.609 – 0.877
—

6b

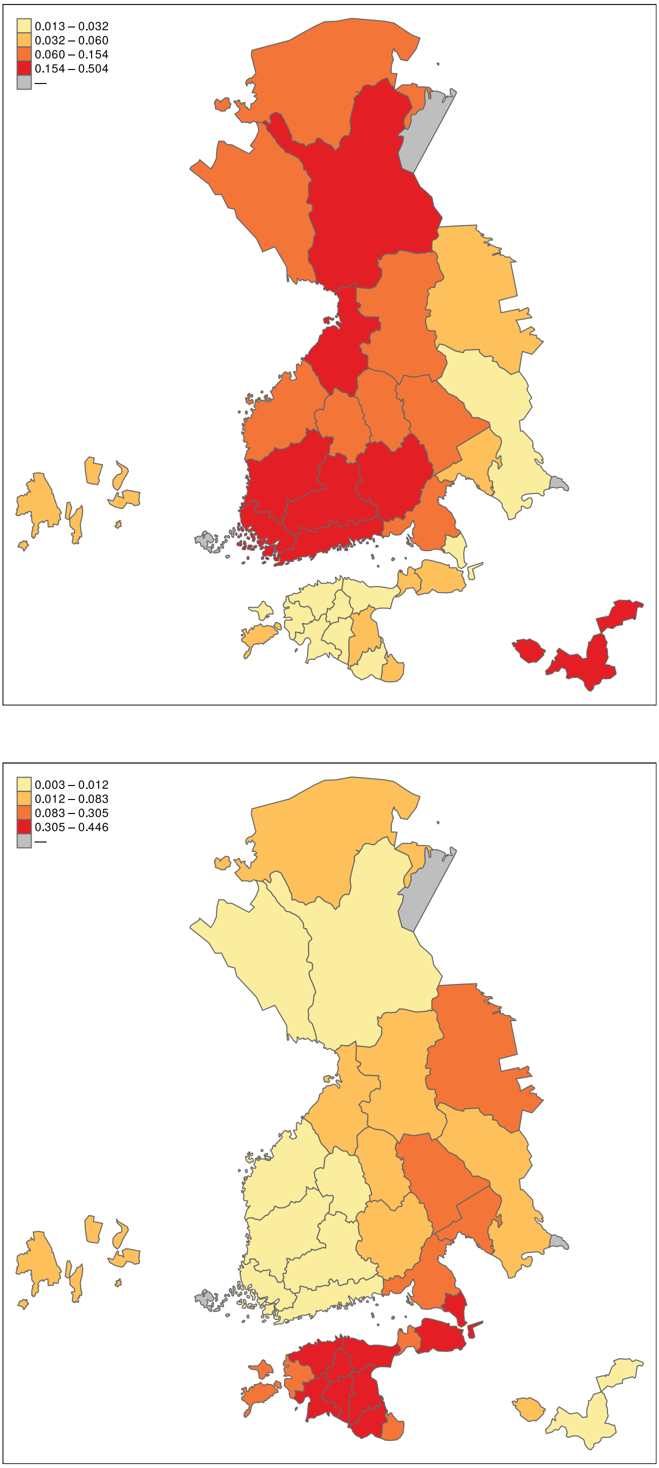


Figure 6. *Shares of the genres by counties: a) narrative songs; b) charms; c) children's songs; d) lyrical songs. The maps are created by combining similar categories with different names in the Estonian and Finnish indices.*

The comparison of geographical distributions of the main categories of the indices gives an overall view of the regional variation of oral cultures and the collecting history. Yet, one category of the index may contain various kinds of poems. For example, charms can have very different characteristics in different runosong regions or within them, ranging from Christian to pre-Christian, from very short to very long, from prose to verse, from a short order or spell to complex negotiations, narratives, and prayer-like texts. Likewise, other categories may contain very heterogenous characteristics. In Figure 6, we give the shares of a) narrative songs, b) charms, c) children's songs, and d) lyric songs in relation to the whole amount of data by counties in the ERAB and SKVR.

Considered by counties, the main genres of the indices, and the whole recording period of the corpora, narrative songs especially dominate the collections from Russian, Tver, and southern Karelia, Karelian Isthmus, and Ingria (Fig. 6a). The share of charms is biggest in western Finland and Scandinavian Finnish-speaking areas. Most of these charms are short compared to the long mythologically dense charms in Karelian-language areas and eastern Finland. In Ingrian and Estonian collections, charms in runosong meter are not as prevalent as in other areas (Fig. 6b). As with charms, the prevalence of children's songs in western and northern Finland is mostly explained by intense collecting activities in the area at the beginning of the twentieth century by students, schoolchildren, and some individual collectors, such as Jenny and Samuli Paulaharju. Unlike other genres, short charms and children's songs were still widely known and available at the time. The number of children's songs, especially from the 1920s and 1930s, is also high in the Estonian archival materials, which have not been added to the ERAB and are thus not visible in the map (Fig. 6c). It is also well known already from earlier research that at the time of recording, the number of lyrical songs and motifs was highest in Estonia, Ingria, and southern parts of Karelia (Fig. 6d; see, e.g., Siikala 1994).

DETECTING COMMON VERSE TYPES IN NORTHERN AND SOUTHERN FINNIC LANGUAGES

The common Finnic database offers possibilities to get a better overview of the whole runosong tradition and to find ways to bring together similar song texts, motifs, and formulas. The evident issue is linguistic variation – the similar verses in different languages and dialects do not match exactly but have several, and in some cases a considerable number of slightly (or more heavily) variable versions. In many cases, it is practically impossible to figure out all the possible dialectal, orthographic, and poetic variations of even one word for text searches. We have, e.g., Väilämöinen, Viänämöinen, Vainämoinen, Wäinämöisen, Väinämöizen, Väinämyösen and over 200 other variants for the old sage Väinämöinen known in the Northern Finnic area.

In order to handle this variability and to bring together similar but slightly variable verse lines in the database, Maciej Janicki developed a method based on bigram vectors, treating verses as if they were words. Our aim was to recognize 'verse types', which we define here as the clusters of such poetic lines that contain the same content words across different dialects but may vary at the surface level and in the use of grammatical words or morphological endings. The method is based on computing the cosine similarity of the lines' character bigram frequencies, and algorithmic clustering based on this similarity measure, and was tested with some common poem types from the SKVR corpus of Northern Finnic languages (Janicki & Kallio & Sarv 2023). The method also often recognizes shorter than line formulas. Yet, the bigram similarity method does not always capture more distant variants of the same verse type. This is not surprising, since the stems, grammatical words and morphological endings in Finnic dialects vary on a regular basis. Between more distant languages – or languages that have been written down in very different orthographies – the similar line level formulas or line types are often not recognized.

In order to detect the commonalities in Northern and Southern Finnic songs computationally, we proceeded with the verse clusters that contained verses from both corpora, altogether 1,844 verse clusters, ordered by size.

The hits at the beginning of the list were part of large clusters, and were mostly not exactly what we were looking for, but clusters consisting of verses from:

1. various sets of similar alliterative or other sound patterns with no content similarity (these may be used further in poetic analysis);
2. refrains, often with many sound repetitions, with potential to help to recognize the refrains in the SKVR corpus where the refrains have not been tagged;
3. a common word or word beginning repetitions, such as "laula laula", "laula laulujani", "laulab laulust" ('sing sing', 'sing songs', 'sings of song') or "kuku kuku", "kukkus kukku", "kuk kuk kuku" ('call, call', 'called call', 'ca-ca-call', often about the sound of the cuckoo), which tell a great deal about the importance of songs, singing, and cuckoos in these traditions. For the scope of this article, these connections are often too vague, not helping in finding closed counterparts;
4. Ingrian poems included in both corpora, sometimes even recorded from the very same singers such as Mikko Pukonen or affected by the folk song publication *Pieni runon-seppä* (Europaeus 1847). The presence of

Ingrian poems in the ERAB is an important observation, but for further experiments in finding the connections between Southern and Northern Finnic language areas – rather than corpora – we exclude these by leaving out the whole small category of Estonian *välismaa* 'foreign country' category (1840 texts);

5. verse types from similar poems and motifs in Southern and Northern Finnic materials.

The main observation from this experiment is that in this type of a computational approach, meaningful similarities at shorter than line and line levels, and poetic or sound repetition similarities are not distinguishable – but can be processed further by close reading of the cases.

Case 5 of actual similar poems and motifs in Southern and Northern Finnic materials offers some really nice results, such as some key verses from the "Singer's Reward" and "The Maiden to Be Ransomed" or from some motifs used in different regional contexts, such as 'In the cloud, there are water drops' used in "Big Oak" in Karelia, in "Making of the Zither" in Ingria, and in "Sacred Grove, Gold Burns" in Estonia, or "It's Good to Be Married to a Crippled One", used in several poem types in Estonia, Ingria, Karelia, and Finland, and also as a short proverb.

Among the closely similar poem type pairs are some interesting loans like the verse "ehittelin, kengittelin" (in poem type "Killer of the Daughters", also used in some other types). In Ingrian and some Estonian versions the verse has a clear meaning of 'putting on nice clothes, putting on shoes', whereas in many Estonian variants the last word has lost its meaning and transformed to something similar by sound but obviously less meaningful (*epitelin*, *kenitelin*, *kehitelin*). Also there are several Estonian–Ingrian verse parallels that in Estonia are known only or mostly on the northern coast which, historically, had closer contacts with Ingria (e.g., *mies vihane vitsikkosta* 'an angry man [came] from the bushes'). Using close reading of the verse pairs, we also detect some rare shared poem types, for example "Mis viga Virus elada" / "Saaressa hyvä elää" 'It's good to live in Viru / It is good to live in Saari', with a bunch of similar lyrical motifs about good life.

Our need to find ways to computationally manage the repeating words, word parts, and sound patterns led to further experimenting with clustering to tune the similarity calculations. In plain calculation of bigrams for verse similarity assessment (*default* method in Runoregi), the sound repetitions – essential to the poetic system of runosong – often bring together the verse lines with similar alliteration patterns instead of similar content. This is useful for studies on poetic and semantic structures and connections created by sound patterns. Yet, as our present interest is to diminish this, we tested with the versions with square root of the count of each bigram (*sqrt*) and just listed the occurrence of different bigrams without counting them (*binary*). Currently, the *default* clustering method often brings false positives, *sqrt* and *binary* clustering less so – but tighter settings may also leave out more potential positives.

A good example of the difference of these methods is the clustering of the verse *tii-tii tihane* ('ti-ti, tit') known mostly in children's songs from Estonia and Finland.

a) The *default* method brings together Estonian and Finnish verses from this poem type, but the repetition of the *ti*-syllable causes also the other verses containing the same alliteration pattern *tii-tii* (*tiiti, tiiti, tengan löysin* 'dingle-dangle, I found money' or *aja tii tiigi poolõ* 'make your way towards the pond') to cluster together with the main verse, altogether 665 cases in the database.

b) The *sqrt*-method clusters together 463 cases of Estonian and Finnish poem types (*tii-tii tihane, tii-tii tianen*) along with some sound repetitions in other types (*ti-ti-tiit, ti-ti-tiit*).

c) The *binary* method is not able to bring together the main Estonian and Finnish versions of the verse, and clusters them separately (with 191 and 135 cases respectively) with plenty of smaller neighbouring clusters of the variations of the same verse type.

In this case, the *sqrt*-method captures the idea of 'verse type' most reliably, but depending on sound patterns, orthographies, and morphological variation, this may be different in the case of other verse types and language areas.

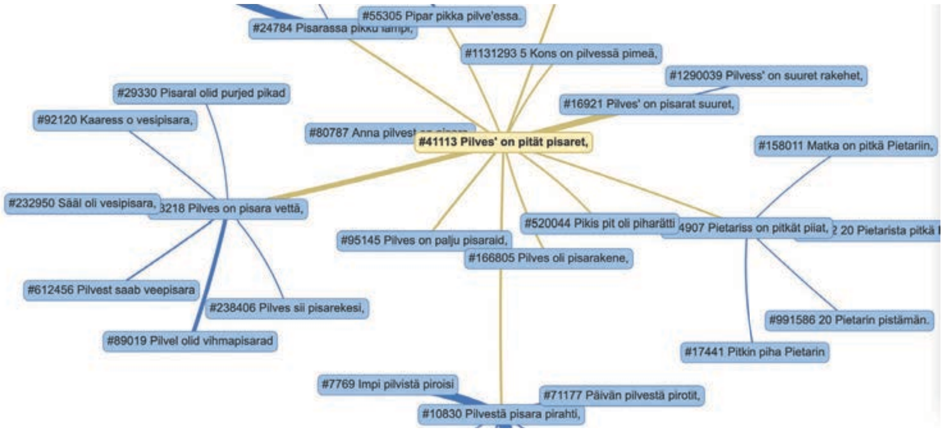

#55305 Pipar pikka pilve'essa.
#24784 Pisarassa pikku lampi,
#1131293 5 Kons on pilvessä pimeä,
#1290039 Pilvess' on suuret rakehet,
#29330 Pisaral olid purjed pikad
#16921 Pilves' on pisarat suuret,
#92120 Kaaress o vesipisara,
#80787 Anna pilvest
#41113 Pilves' on pität pisaret,
#158011 Matka on pitkä Pietariin,
#232950 Sääl oli vesipisara,
3218 Pilves on pisara vettä,
#520044 Pikis pit oli piharätti
4907 Pietariss on pitkät piiat,
2 20 Pietarista pitkä
#95145 Pilves on palju pisaraid,
#166805 Pilves oli pisarakene,
#612456 Pilvest saab veepisara
#238406 Pilves sii pisarekesi,
#991586 20 Pietarin pistämän.
#89019 Pilvel olid vihmapisarad
#17441 Pitkin piha Pietarin
#7769 Impi pilvistä piroisi
#71177 Päivän pilvestä pirotit,
#10830 Pilvestä pisara pirahti,

Figure 7. *Network of nearby verse clusters for the cluster that contains the line "Pilves' on pität pisaret" ('There are long water drops in the cloud'), characteristic of the poem type "Big Oak" in Karelia and other song types in Ingria and Estonia.*

Experimenting also led to various developments of the similarity recognition interface Runoregi to help with close reading and assessing the computational results. A relevant addition in terms of the present paper, resulting from the notion that the regional verse type versions often get into nearby clusters, was to add a neighbouring verse clusters view (Fig. 7). This helps the users to quickly estimate the relations, coverage, and quality of different clusters.

LOWER VERSE SIMILARITY THRESHOLD OF SEVERAL VERSES

From earlier research and our manual experience with the corpora we know that the list of similar verse pairs presented above does not cover all the relevant verse type or motif similarities between Northern and Southern Finnic languages. Thus, here we develop the method further, using a low similarity threshold for verse similarity detection (with *default* method), demanding several lines of a poem pair to indicate some similarity, and grouping these similar poem pairs by corresponding poem types, one from the Estonian and the other one from the Finnish index. The similarity threshold is so low that, as such, it would produce more false positives than is feasible to sort out manually. Yet, requiring several slightly similar verses per poem pair raises the probability of finding relevant cases. Here, we first describe the method and then take a look at the top 20 results.

Using the FAISS library for efficient similarity search in sets of vectors (Johnson & Douze & Jégou 2017), we were able to identify pairs of lines with a similarity higher than 0.75 across the entire corpus, which was enough for identifying lines with the same content. The similarity measure could also be used to compute linewise side-by-side alignments of texts using the weighted edit distance algorithm (Janicki 2022).

Experiments with alignments have shown that the same character bigram-based similarity measure could be used for recognizing commonalities crossing the Northern-Southern Finnic boundary, if only the similarity threshold could be lowered to around 0.5 (see Table 1). A looser threshold would allow the detection of similar verses despite language differences (for example in morphology) as the shared verse types are often similar not only in meaning, but also in sound patterns, using cognate words and similar metrical structure. For these reasons, when we experimented with machine translation, it turned out to be of little use as it does not preserve euphonic features and does not manage the non-standard language sufficiently.

The similarity value of 0.5 is still higher than the average for a random pair of lines, but many unrelated lines can achieve it due to coincidence and similarity of sound patterns only. Thus, finding and listing all verse pairs with such similarity did not seem feasible. Instead, the similarity recognition needed to aim at poems containing several such line pairs in a sequence. A model example of this was a set of versions of the song "The Maiden to Be Ransomed", a well-known poem in a transitive form between the old runosongs and stanzaic songs, which follows an unusually fixed structure preserved across languages (Table 1; see Kemppinen 1957).

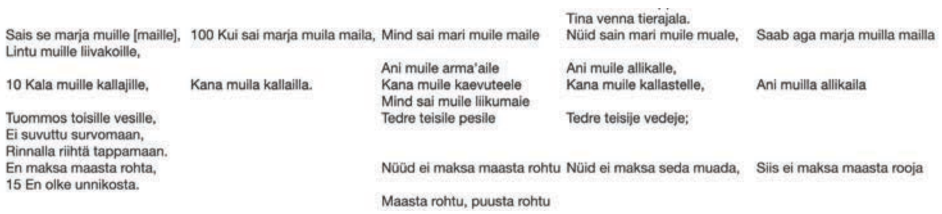
Table 1. *Fragment of an Ingrian-Finnish and Estonian version of the song "The Maiden to Be Ransomed", showing the possibility of cross-lingual alignment. The similarity values above 0.5 are shown in bold. Based on the vocabulary (aik' oli ikäv, reissivanna, vällaa), this Ingrian version is probably translated quite directly from Estonian, which partly adds up to the similarity.*

Ingrian-Finnish	**Estonian**	**translation**	**sim.**
Lilla istu kamperissa,	Lilla istus kammeris,	The girl was sitting in a chamber,	**.79**
Aik' oli ikäv uottaa,	Tal aeg oli igav oota.	It was a sad time waiting.	.46
Näki vennan reissivanna	Ta nägi venda sõudema	She saw a brother [travelling / rowing]	.20
Pitkin mere rantaa.	Seal üle mereranna.	Along the seacoast.	.45
"Rikas venna, rakas venna,	"Kulla venda, rikas venda	'Rich brother, [dear / golden] brother	**.64**
Lunast minnuu täältä vällää!"	Lunasta mu südant!"	Ransom [me / my heart]!'	.31
"Millä mie lunassan,	"Kellega ma lunastan,	'How can I ransom you,	.41
Kui miull' ei ole varraa?"	Kui mul ei ole raha."	If I don't have money?'	**.73**
"On siull' koton kolme miekkaa,	"Sul on kodu kolmi mõeka,	'You've got three swords at home,	**.66**
Pane niist' yksi pantiks!"	Pane üks neist pandiks."	Pawn one of them!'	**.74**
"Enne mie luovun siusta	"Ennem mina lahkun õekesest,	'I'd rather give up [you / a sister],	.36
Kui omast' kolmest' miekast'."	Kui oma sõjamõegast."	Than my own [three / war] sword[s].'	.44

A breakthrough was achieved by optimizing the weighted edit distance algorithm to work with many poem pairs simultaneously (Janicki 2022, 2023), which allowed for computing the alignment between all poem pairs in the corpus, using any similarity threshold. The decision on whether to keep a poem pair as a result could thus be made based on its alignment. Right now we simply put a threshold on the total similarity of aligned lines, but in the future, the possibility to take into account their distribution in the poems (roughly contiguous passages vs. single scattered lines) will be studied as well.

We have deliberately set thresholds to low values so that as many similarities as possible could be captured. Due to that, the precision of the method is low, i.e., many pairs are in fact unrelated (false positives) and only look similar due to similar sound patterns or coincidence. Thus, the results need to be verified by close reading and cannot readily be used for quantitative summaries. The planned future work on analysing the distribution patterns of similar lines in the texts might help filter out the false positives computationally.

The computation produced over 41 million poem pairs recognized as containing sufficient similarity, of which there are 55,910 pairs of a poem from the SKVR and a poem from the ERAB. To facilitate the analysis of the results, they are grouped by type names according to SKVR and ERAB type indices. Each group relating to a particular pair of poem types in the Estonian and Finnish index may thus contain different kinds of similar verse pairs. The user interface Runoregi is used to browse the examples as either pairwise or multiple-sequence alignments, which helps to spot the passage that the similarity is based on (Fig. 8). We also generated a map link to quickly see the areal distributions of grouped poems.



Sais se marja muille [maille], Lintu muille liivakoille,	100 Kui sai marja muila maila,	Mind sai mari muile maile	Tina venna tierajala. Nüid sain mari muile muale,	Saab aga marja muilla mailla
10 Kala muille kallajille,	Kana muila kallailla.	Ani muile arma'aile Kana muile kaevuteele Mind sai muile liikumaie	Ani muile allikalle, Kana muile kallastelle,	Ani muilla allikaila
Tuommos toisille vesille, Ei suvuttu survomaan, Rinnalla riihtä tappamaan.		Tedre teisile pesile	Tedre teisije vedeje;	
En maksa maasta rohta, 15 En olke unnikosta.		Nüüd ei maksa maasta rohtu	Nüid ei maksa seda muada,	Siis ei maksa maasta rooja
		Maasta rohtu, puusta rohtu		

Figure 8. *Fragment of a multiple-sequence alignment of some similar texts from Western Ingria (Narvusi and Kattila) and Northern Estonia (Virumaa and Harjumaa) in the poem type "At Paternal Home and at Husband's Home". Detail of a larger view generated by Runoregi.*

To help in close reading, the table also includes links to the poem types in Runoregi, to a map projection and to a set of 2–15 most similar texts in an aligned view. The present table has 7,502 rows and includes similarities of 1,838 different song types and 13,295 individual texts. Thus far, we have checked a few dozen most similar type-to-type relationships in more detail (with text and metadata search interface Octavo and Runoregi functionalities) and done random checks to different parts of the table. Table 2 gives a short description of the first 20 cases.

Table 2. *First 20 rows of the low threshold similarity calculation by poems organized by poem types.*

	Finnish type name	**SKVR**	**Estonian type name**	**ERAB**	**Connecting feature**
1	Kips kilo karjaan	21	Kits kile karja!	229	Children's chain song[11] of shared origin (Goat go to the herd)
2	Tiiri liiri linnun poika (…)	30	Tii, tii, tihane	34	Common starting formula in some children's songs ("ti ti tit")
3	Kello yks – muna kyps	77	Kell üks	25	Children's rhyme with counting the hours, details vary (One o'clock, the egg is cooked)
4	Laulajan palkka	80	Laulikule palka!	58	Overlap in short motif: "I am not singing without the wage" (Singer's reward)
5	Tiiri liiri linnun poika (…)	69	Liiri-lõõri!	78	Overlapping elements of chain song of shared origin with onomatopoetic type name
6	Avioitumisajan tiedustelu käeltä	9	Kägu kukub	89	One similar line ("call call cuckoo") in two poem types about the cuckoo
7	Lunastettava neito	94	Lunastatav neiu	40	Very similar poems also in structure, probably rather recent spread (The Maiden to Be Ransomed)
8	Tanssituvan pyyntö	23	Ori lahkub	52	A common verse pair addressing the host and hostess used in different poetic contexts (here in these two poem types)

9	Lauloin ennen lapsempana	27	Laulikule palka!	52	Overlap in some verses about singing and a short motif about rewarding the singer
10	Lapsen maitohampaan (…)	38	Hambasõnad	4	Short children's spell about changing a tooth of shared origin (Words of the tooth)
11	Pakeneva	5	Kits kile karja!	175	Two different poem types with some common chain song elements
12	Eliniän tiedustelu käeltä	14	Kägu kukub	85	One similar line "call call cuckoo" in two different poem types
13	Tanssituvan pyyntö	25	Orja palk	54	A common verse pair addressing the host and hostess, two different poem types
14	Käeltä pyydetään onnea (…)	4	Kägu kukub	51	One similar line "call call cuckoo" in two different poem types
15	Varas vie koristeet	78	Hobune varastatud	18	Similar song structure, where a theft of jewellery or a horse happens, and the story will be re-told to the parents who comfort the poetic I: shared elements and formulas, but different poem types
16	Laulan lapselleni (…)	5	Suude sulg	167	Verse variations containing alliterative word pair about singing and child, two different poem types
17	Tii tii tiainen	1	Tii, tii, tihane	34	Shared starting formula of a children's song about birds
18	Kiletoivirsi	22	Ori lahkub	35	A common verse pair addressing the host and hostess used in different poetic contexts (here in these two poem types)
19	Kolme käkeä	20	Kägu kukub	26	One similar line "call call cuckoo" in two different poem types

20	Käärmeen sanat	98	Ussisõnad	17	Very similar variations of snake charm all over the Finnic area

The table includes very different types of similarity, geographical spread, genres, and styles. Similarity may occur at the level of the whole poem types or some elements of them, or a motif or only one shared verse type. The most recurrent cases in the first twenty rows are short songs or sayings for children with two or several similar lines, including one wider set of interlinking poem types ("Tiiri liiri linnun poika", "Tii tii tihane", "Liiri-lõõri", "Tii tii tiainen...") with examples also further down the table. Similar cases are also short 2–3-line formulas about rewarding the singer or greeting the host and the hostess, characteristic of particular poem types but also used elsewhere. "The Maiden to Be Ransomed" presented already above makes a special case in terms of structural similarity of the versions. The song groups sharing one-line formulas about the cuckoo or singing often contain plenty of false positives due to word and sound repetitions, but also some shared line types. Two rows present two different narrative song types that yet share similar motifs, structures, or parts. The "Snake Charm" of the last row exhibits some surprising similarities across the whole runosong area.

Although the Gulf of Finland and the waterways were important contact routes, it is natural that most sets of similar line types occur between regions located next to each other, especially between Virumaa parish of Estonia and Narvusi parish of western Ingria. In our table of recognized poem or line-level connections between the Estonian and Finnish corpus, the poems from Western Ingria dominate (1,996 cases), and southern Karelia (1,356) and Virumaa (1,027) also provide over a thousand texts each. The fewest similarities to the Estonian corpus are found in the most faraway places, in Länsipohja/Västerbotten (5), and in the seventeenth-century migration areas of Tver Karelia (13) near Moscow, and Finnish speaking Värmlanti (25) in central Sweden. Also the Olonets Karelia (83) is underrepresented. Part of this basic setting may relate to the differences in orthography, part to linguistic or cultural differences, and part to the relatively small recordings made from these areas.

Most of the texts that are detected as similar also contain false positives due to our chosen low similarity threshold. It is possible that further experimenting with similarity methods and thresholds to reduce false positive verse pairs helps to sift the most relevant results. Further, if one poem text is given several type names, the similarity results are given for each of these types, making some results repeat.[12] The first complete false positive case comes already at row 22 of the results, connecting many Estonian and one southern Karelian poem based on word and sound repetition relating to cuckoo and singing, with no shared verses proper. Yet, also the very end of the table still contains some relevant results. The last relevant case, a closely related formulaic line pair (“there comes a big eater, a big eater, a big drinker”) in one Estonian and one Ingrian lyrical poem, appears at the rows 7433–7449.

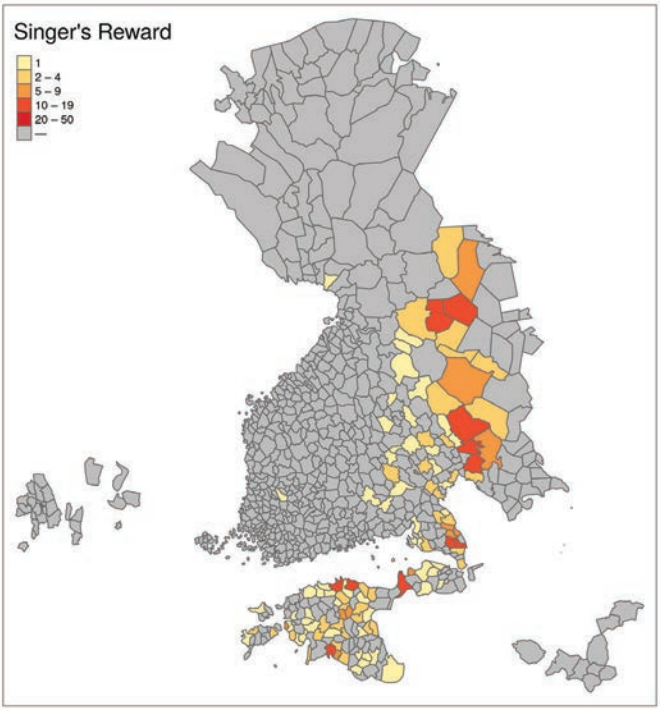
COMPARING DIFFERENT METHODS: SINGER’S REWARD

Due to the great and multilevel variation of oral poetry, different methods available for recognizing similarity in the corpus of Finnic oral poetry yield different reaches, and not everything is yet computationally recognizable. Although methods also function variously with different kinds of verse types, formulas, motifs and poem types, we approach this variability here via one example from the cases above. The one-line formula “I won’t sing without the wage”, often added with “use my mouth without gold” is indexical of the poem type “Singer’s Reward”, and found in all the language areas of the corpus except for the less documented and endangered Votic and Ludic.

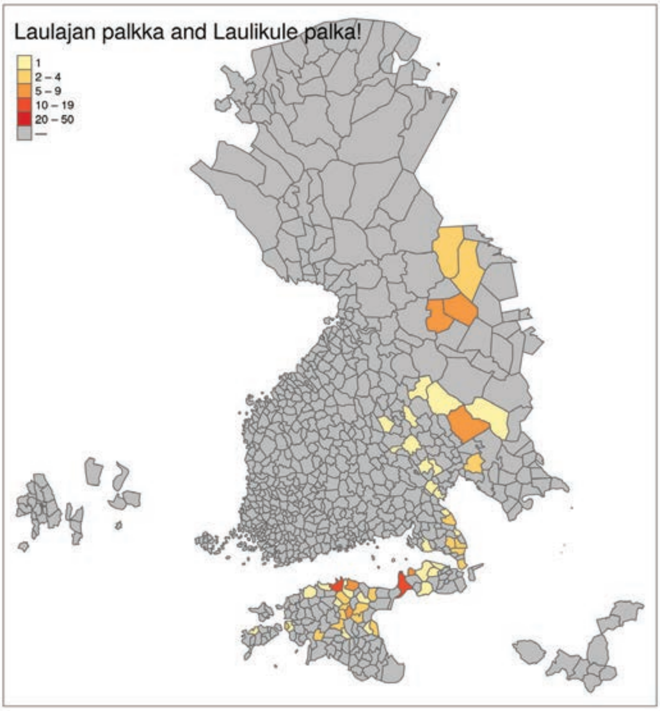
Within the scope of available methods, the manually created poem-type indices typically cover the widest geographical spread and include the greatest number of the material as is the case also here (Fig. 9a). This is natural: the type covers much more than the key formulas, line types, or motifs – not all the similar poems include all the key formulas. The cases in our similarity recognition table have surprisingly similar spread (Fig. 9b) with the simple manually created formula search of the most stable part of the key verse “sing without money” (Fig. 9c): both are narrower than the spread of the poem type indices. Comparison with the relatively rare key formulas that have *not* been indexed to the “Singer’s Reward” (Fig. 9d) indicates that in the indices, the formula is probably understood to strongly index the poem type, and that the formula does not often appear in other kinds of poetic contexts. Yet, not all the texts indexed to the type include the most evident forms of the formula (e): similar ideas about rewarding the singer can be expressed in different ways, and some texts are only short fragments.

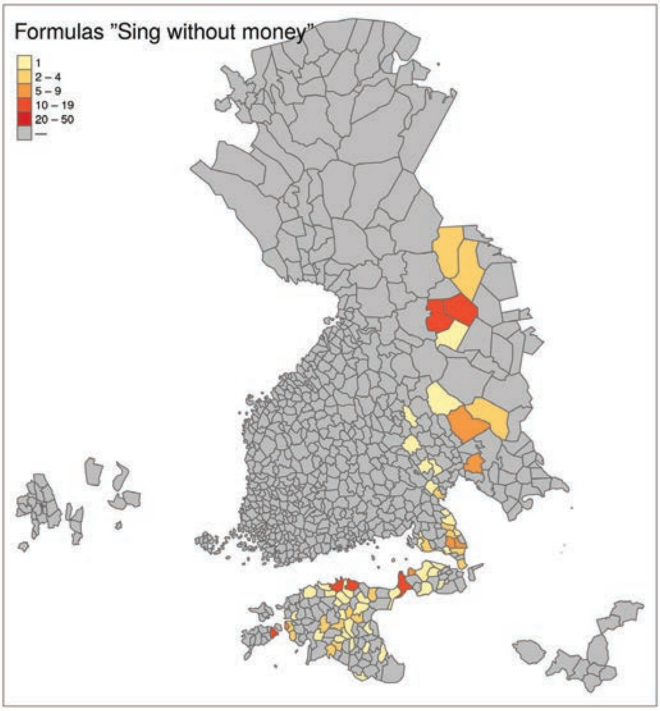
In this case and in many others, the manually created poem type index gives the widest reach. In the Finnish poem type index, the lyrics have been indexed at a detailed level closer to shorter motifs than poem types. Yet, even here, in terms of the verse similarity, the poem type is not defined by the key motif(s) at the level of the line types only, but of the key content. The aim has apparently been to include all the poems mentioning the singer’s reward. Thus, the key motif may be “I won’t sing without the wage / gold”, or, for example, “I am not asking

9a

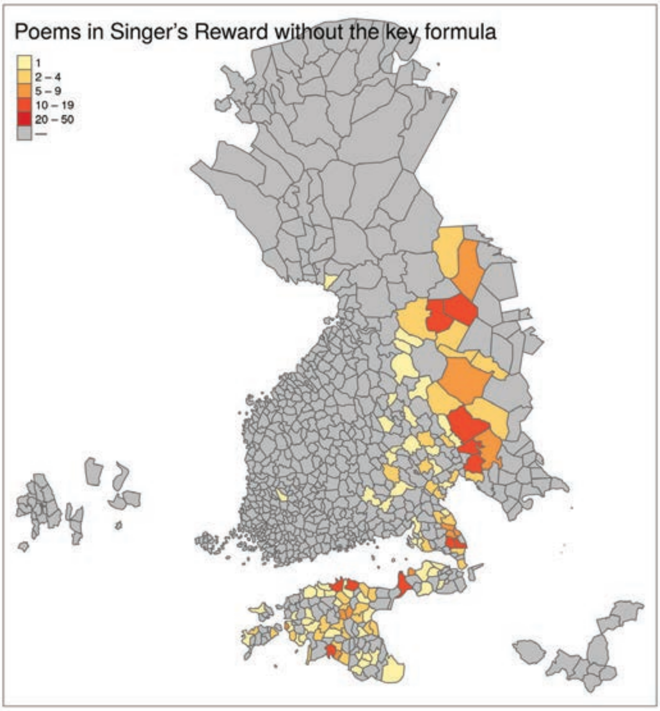

Singer's Reward
1
2 – 4
5 – 9
10 – 19
20 – 50
—

9b


Laulajan palkka and Laulikule palka!
1
2 – 4
5 – 9
10 – 19
20 – 50
—


Formulas ”Sing without money”
1
2 – 4
5 – 9
10 – 19
20 – 50
—

9c


Poems in Singer’s Reward without the key formula
1
2 – 4
5 – 9
10 – 19
20 – 50
—

9d

9e

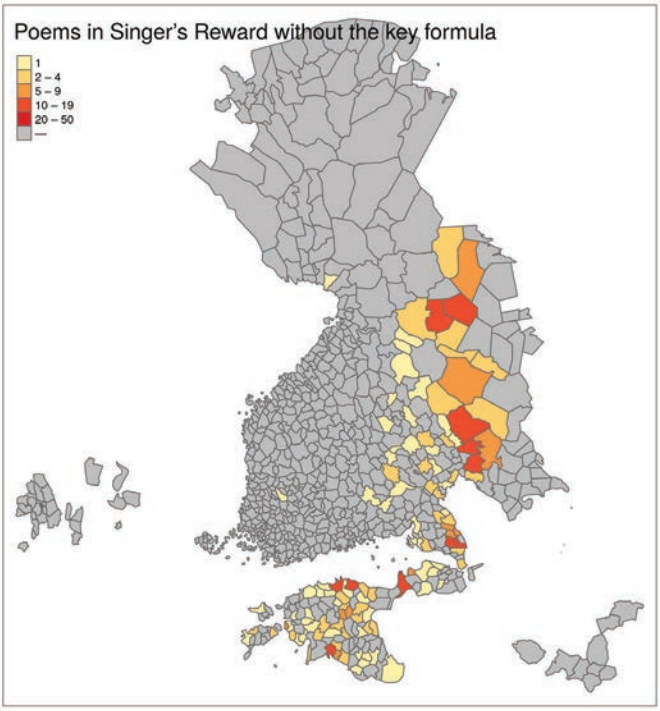
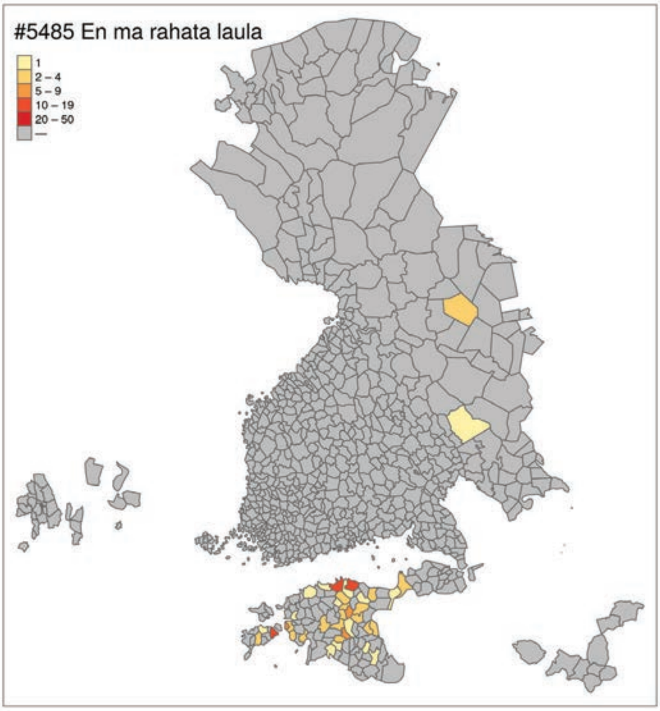


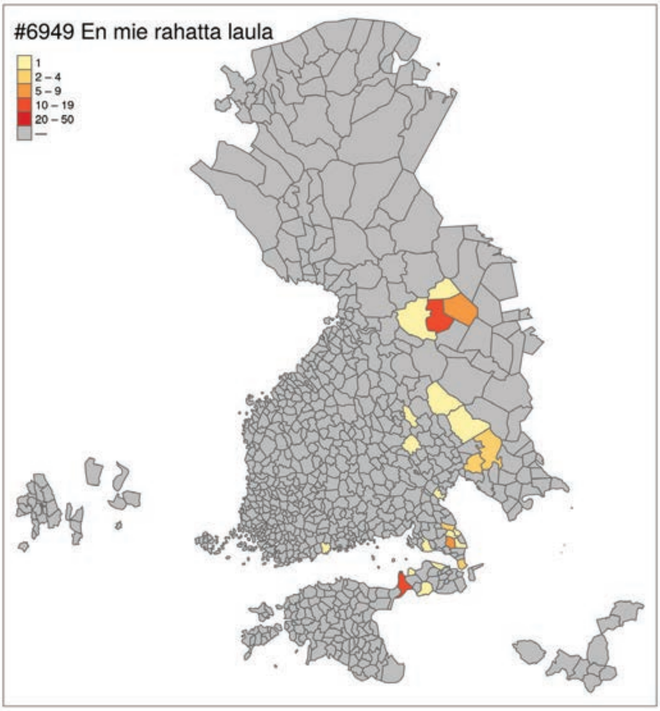
Figure 9. *a) The poems indexed to the poem type "Singer's Reward" (Est. & Fi.), b) poems recognized by the similarity calculation (row 4); c) the formulas "sing without money" (manually made query, false positives excluded);*[13] *d) formulas "sing without money" that are not indexed into "Singer's Reward"; e) "Singer's Reward" poem types which do not include the most evident forms of the formula "sing without money".*

much: one [Russian] kopeck for a whole word [=line], half for half a word, one öre [Swedish penny] for a movement of the tongue", or "My tongue won't sing without festive pies, my mind without black mead, my lips without hop water".

The current Runoregi interface applies a higher similarity threshold than our experiment in this paper, and often does not reveal similarities between the Estonian and Finnish collections, or between more distant languages or dialects. It is also typical for the interface that one line type can – due to dialectal, poetic, and orthographic variation – distribute into several line clusters (see Fig. 10) with different regional spreads. Interestingly, the biggest three Runoregi clusters of the line type "I won't sing without the wage" have wide and partly overlapping spreads, and the first of them even connects line types in the Estonian and Finnish corpora. It looks like the line type is not only stable and widespread but also has such a bigram structure that it gets recognized easily and, computationally, does not merge with lines with other content.


#5485 En ma rahata laula
1
2 – 4
5 – 9
10 – 19
20 – 50
—

10a


#6949 En mie rahatta laula
1
2 – 4
5 – 9
10 – 19
20 – 50
—

10b

10c

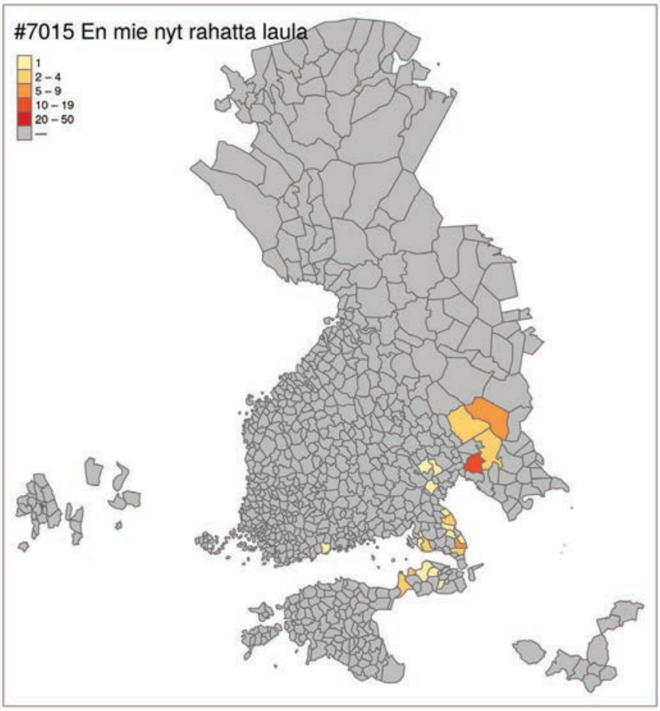


Figure 10. *Three biggest line clusters (default clustering) in Runoregi of different sub-cases of the line type "I won't sing without the wage". Clusters contain different variations of the line type. These maps also include cases from the JR corpus.*

CONCLUSIONS

The variation of oral poetry is complex and takes place at various levels of content and form. This means that no single computational method or toolbox is able to track all the aspects of it, especially in a corpus that represents small non-standardized dialectal language variants.

It is evident that only part of the Finnic data has been treated in earlier scholarly analysis. Particular genres, poem types, and regions have attracted most of the attention, often treated in separate studies. Thus, we do not know how common and significant the different patterns of variation at the level of the whole data are, and cannot be sure that the research has noticed everything that is relevant in terms of less studied parts of the data. This means that the data-driven approaches have a true potential to reveal new things, especially if combining computational or quantitative methods with close reading and qualitative interpretation.

In this article, our approach has been to explore the similarities of two corpora in related languages computationally, also close reading of the results, projecting them to a map, and comparing the results to one another and to the manually created poem type indices. The similarity calculation methods presented here help to recognize connections in the wide corpus of oral poetry and are relevant also across the linguistic and cultural border of Estonian and Northern Finnic languages (Karelian, Ingrian and Finnish). The results are noisy and need human interpretation to sort out the relevant cases. The methods may also be used for purposes other than the ones in this article: our first experiments bring up many cases of similarity at the level of sound repetitions (e.g., alliterative patterns and refrains) rather than verse types or content.

Yet, based on earlier research and on our own manual comparison of some poem types in the ERAB and SKVR indices, it is also evident that our similarity recognition method – even the fine-tuned one – does not catch all the similarities. In some cases, this is due to different phonetic forms in different related languages resulting in too different word forms and too different bigram structures, even if the human reader may recognize the lines to be very similar in form and content. In some cases, this same happens because the two versions share similar content but are told with different words and formulas (see, e.g., cases in Krohn 1931). Computational recognition of texts with similar content but totally different word stems and formulas would necessitate tools that are not yet available for our historical poetic and dialectal data with plenty of variations.

It is not a trivial task to bring together the shared features in different linguistic areas computationally. In our experiments, especially the borderline between Southern and Northern Finnic languages has been difficult to cross, although it is yet to be analysed how much we are able to recognize similarities of data, e.g., in Votic or Livvi (Olonets) Karelian languages in relation to other Finnic areas. Here, the problems arise from intertwined linguistic (morphological, lexical, and orthographic) and poetic or cultural differences. Words, formulas, line types, motifs, and poem types may take different forms and be used in different ways and contexts, which is in line with what has been thought of the variation of regional song cultures, e.g., by Anna-Leena Siikala (1994). Typically, both the content and form of poems are the more distant the more there is linguistic or geographical distance between the instances (see Sarv & Kallio & Janicki 2024). Yet, the cases revealed in our exploration, such as

"The Maiden to Be Ransomed" or "Snake Charm", signal that there may be interesting exceptions to this.

ACKNOWLEDGEMENTS

The research presented in this article was conducted within the projects PRG1288 of the Estonian Research Council; TK215 and EKKD-III2 of the Ministry of Education and Research, projects 333138, 333139 and 346342 of the Research Council of Finland, and the project "Itämerensuomalaisen kansanrunouden muodon ja sisällön variaatio suhteessa kielelliseen ja etniseen historian" (Variation of content and form of Finnic oral poetry in relation to linguistic and ethnic histories) funded by the Kone Foundation.

NOTES

[1] *Regilaul* 'regi-song' in Estonian (< *Reigenlied* 'dancing song' in German), *runolaulu* (runo-song, poem-song) in Finnish and Karelian (< *rūnō* 'secret', 'magic sign', 'knowledge', 'wisdom' in Proto-Germanic) or, traditionally, just *laul*, *laulu* (song), or *virsi, vers* ('poem', nowadays 'Lutheran hymn') in Finland, Ingria, and Karelia, or descriptively, "old folk poems", "Kalevalaic poems" or "poems in Finnic alliterative tetrametric trochee" (see Kallio & Frog & Sarv 2017).

[2] For discussions on Kalevipoeg as an epithet or character in Finnic oral poetry and as a main character especially in prose stories, see, e.g., *Elias Lönnrotin kirjeenvaihto*: Elias Lönnrot > Emil Sachsendahl, 4.1.1851; Friedrich Reinhold Kreutzwald > Elias Lönnrot, 20.2.1851; Krohn 1903–1910).

[3] See Kundozerova 2022 about the situation in Russian Karelia.

[4] "Formulaic intertextuality, thematic networks and poetic variation across regional cultures of Finnic oral poetry", funded by the Research Council of Finland, 2020–2024, at the Finnish Literature Society and University of Helsinki, in close collaboration with the Estonian Folklore Archives; see https://blogs.helsinki.fi/filter-project/, last accessed on 21 October 2024.

[5] Structured Query Language.

[6] Runoregi is currently available at https://runoregi.rahtiapp.fi/ (last accessed on 21 October 2024), and will be maintained at least for some years after the project. The similarity calculation method, codes and computational pipeline for setting up Runoregi with FILTER or other data are published on GitHub (Janicki 2024a, 2024b).

[7] It is also supplemented by some literary works in runosong meter, which are not analysed in this article.

[8] Optical Character Recognition.

[9] On the one hand the high number of types with a single text is normal; Arvo Krikmann has shown that Zipf's law known in linguistics, which describes the distribution of word use in text (very few very frequent words, very many very rare words), also applies to the type distribution of archival collections of various folklore genres (Krikmann 1997). On the other hand, part of the variation may be attributed to the variability of type names rather than types due to the decades-long copying process.

[10] The coloured version of the figure is available in the online publication.

[11] Chain song means the poem mostly builds on interlinked verses where the lines repeat one word or word pair of the preceding line (see Saarinen & Janicki & Kallio forthcoming).

[12] Although we have experimented with this, we are not able to reliably sort the line types by poem types. This is caused by both the complexity of oral poetry, where one line type can be used in different contexts and sometimes even in different meanings, and the heterogenous and partly subjective character of the type indices.

[13] Octavo query: ("raha* laula*" OR "raho* laula*" OR "laul* raha*") (collection:skvr OR collection:erab) -(type_id:erab_002001041 OR type_id:skvr_t020100_1040) –"võerad ju omasse" -(e000197980000 OR e000848800000 OR kvr13040160 OR h23201240022 OR ve0301800000 OR h33002960000 OR h20908090095 OR e000848800000 OR e000555340011) -rahasäk*.

SOURCES, INTERFACES, AND CORPORA

Elias Lönnrotin kirjeenvaihto-verkkoaineisto. [Correspondence of Elias Lönnrot, Online Publication.] 2017–. Finnish Literature Society. Available at http://lonnrot.finlit.fi/omeka, last accessed on 21 October 2024.

ERAB = *Eesti Regilaulude Andmebaas* [Database of Estonian Runosongs]. 2003–. Compiled by Janika Oras & Liina Saarlo & Mari Sarv & Kanni Labi & Merli Uus & Reda Šmitaite. Tartu: Estonian Folklore Archives of the Estonian Literary Museum. Available at https://www.folklore.ee/regilaul/andmebaas. (Version of April 2023.)

Europaeus, David Emanuel Daniel 1847. *Pieni runon-seppä eli Kokous paraimmista Inkerinmaan puolelta kerätyistä runo-lauluista.* [Small Smith of Songs. A Collection of Runo-Songs Collected from Ingria.] Helsinki: J. Simeliuksen perilliset. Available at https://urn.fi/urn:nbn:fi:sks-dor-002039, last accessed on 21 October 2024.

FILTER database 2020–. Compiled by Maciej Janicki & Eetu Mäkelä with FILTER project team. University of Helsinki & Finnish Literature Society & Estonian Literary Museum. See description at https://github.com/hsci-r/filter-pipeline, last accessed on 6 November 2024.

Janicki, Maciej 2024a. Runoregi. Script for setting up Runoregi user interface for Finnic oral poetry. *Github*. Available at https://github.com/hsci-r/runoregi, last accessed on 21 October 2024.

Janicki, Maciej 2024b. FILTER Pipeline. Scripts for creating the corpus and database used in the FILTER project. *GitHub*. Available at https://github.com/hsci-r/filter-pipeline, last accessed on 21 October 2024.

JR corpus = *Julkaisemattomat runot.* [Unpublished Poems.] Digitized version, February 2024. Helsinki: Finnish Literature Society.

Octavo UI = *Finnic Oral Poetry.* 2017–. Created by Eetu Mäkelä. Available at https://jiemakel.github.io/octavo-nui/#/search? endpoint=https%3A%2F%2Ffilter-octavo.rahtiapp.fi%2Ffilter%2F&level=POEM, last accessed on 21 October 2024. University of Helsinki.

Runoregi 2022–. Created by Maciej Janicki & Kati Kallio & Mari Sarv & Eetu Mäkelä. Helsinki & Tartu: University of Helsinki (HELDIG) & Finnish Literature Society & Estonian Folklore Archives. Available at https://runoregi.fi. (Version from 15.01.2024.)

SKSÄ 2023:3. Interview with Senni Timonen, 13.2.2023, Helsinki. Interviewers Jukka Saarinen and Kati Kallio. Archives of the Finnish Literature Society, Collection of Traditional and Contemporary Culture.

SKSÄ 2023:30. Interview with Senni Timonen, 23.2.2023, Helsinki. Interviewers Jukka Saarinen and Kati Kallio. Archives of the Finnish Literature Society, Collection of Traditional and Contemporary Culture.

SKVR corpus. Version 2.0. (13.9.2022). Finnish Literature Society. Available at https://github.com/sks190/SKVR, last accessed on 21 October 2024.

SKVR online service 2004–. Edited by Jukka Saarinen & Arvo Krikmann. Helsinki: Finnish Literature Society. URN:[NBN:fi-fe20051411.] Available at https://skvr.fi, last accessed on 21 October 2024.

REFERENCES

Abello, James & Broadwell, Peter M. & Tangherlini, Timothy R. & Zhang, Haoyang 2023. Disentangling the Folklore Hairball: A Network Approach to the Characterization of a Large Folktale Corpus. *Fabula*, Vol. 64, No. 1–2, pp. 64–91. https://doi.org/10.1515/fabula-2023-0004.

Eklund, Johan & Hagedorn, Josh & Darányi, Sándor 2023. Teaching Tale Types to a Computer: A First Experiment with the Annotated Folktales Collection. *Fabula*, Vol. 64, No. 1–2, pp. 92–106. https://doi.org/10.1515/fabula-2023-0005.

Frog 2019. The Finnic Tetrameter – A Creolization of Poetic Form? *Studia Metrica et Poetica*, Vol. 6, No. 1, pp. 20–78. https://doi.org/10.12697/smp.2019.6.1.02.

Frog 2021. "Suomalainen koulukunta": Suomalainen folkloristiikka metodisten jatkuvuuksien ja muuttuvien paradigmojen välillä. [Finnish School: Finnish Folklore Studies between Methodological Continuums and Changing Paradigms.] In: Niina Hämäläinen & Petja Kauppi (eds.) *Paradigma: Näkökulmia tieteen periaatteisiin ja käsityksiin*. Helsinki: SKS, pp. 59–88. https://doi.org/10.21435/ksvk.100.

Grünthal, Riho 2020. The Spread Zones and Contacts of Medieval Finnic in the Northeastern Baltic Sea Area: Implications for the Rate of Language Change. *Journal of Historical Sociolinguistics*, Vol. 6, No. 2, 20190029. https://doi.org/10.1515/jhsl-2019-0029.

Harvilahti, Lauri 2013. The SKVR Database of Ancient Poems of the Finnish People in Kalevala Meter and the Semantic Kalevala. *Oral Tradition*, Vol. 28, No. 2, pp. 223–232. http://dx.doi.org/10.1353/ort.2013.0019.

Harvilahti, Lauri 2019. History of Computational Folkloristics in Finland and Some Current Perspectives. In: Pekka Hakamies & Anne Heimo (eds.) *Folkloristics in the Digital Age*. Helsinki: Suomalainen Tiedeakatemia, pp. 158–175.

Hautala, Jouko 1954. *Suomalainen kansanrunoudentutkimus.* [Finnish Folklore Studies.] Helsinki: Suomalaisen Kirjallisuuden Seura.

Ilyefalvi, Emese 2020. Distant Reading of the Metadata of the Digitized Hungarian Charm Corpus. *Incantatio*, Vol. 9, pp. 113–132. https://doi.org/10.7592/Incantatio2020_9_Ilyefalvi.

Jänicke, Stefan & Wrisley, David Joseph 2017. Visualizing *Mouvance*: Toward a Visual Analysis of Variant Medieval Text Traditions. *Digital Scholarship in the Humanities*, Vol. 32, Suppl. 2, pp. ii106–ii123. https://doi.org/10.1093/llc/fqx033.

Janicki, Maciej 2022. Optimizing the Weighted Sequence Alignment Algorithm for Large-Scale Text Similarity Computation. In: M. Hämäläinen & K. Alnajjar & N. Partanen & J. Rueter (eds.) *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities.* Taipei: Association for Computational Linguistics, pp. 96–100. Available at https://aclanthology.org/2022.nlp4dh-1.13, last accessed on 21 October 2024.

Janicki, Maciej 2023. Large-Scale Weighted Sequence Alignment for the Study of Intertextuality in Finnic Oral Folk Poetry. *Journal of Data Mining and Digital Humanities*, NLP4DH. https://doi.org/10.46298/jdmdh.11390.

Janicki, Maciej & Kallio, Kati & Sarv, Mari 2023. Exploring Finnic Written Oral Folk Poetry through String Similarity. *Digital Scholarship in the Humanities*, Vol. 38, No. 1, pp. 180–194. https://doi.org/10.1093/llc/fqac034.

Janicki, Maciej & Kallio, Kati & Sarv, Mari & Mäkelä, Eetu 2024. Runoregi: A User Interface for Exploring Text Similarity in Oral Poetry. *Proceedings of the 8th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2024).* https://doi.org/10.5617/dhnbpub.11523.

Johnson, Jeff & Douze, Matthijs & Jégou, Hervé 2017. Billion-Scale Similarity Search with GPUs. *arXiv:1702.08734*. https://doi.org/10.48550/arXiv.1702.08734.

Kalkun, Andreas 2015. *Seto laul eesti folkloristika ajaloos: Lisandusi representatsiooniloole.* [Seto Songs in the History of Estonian Folklore Studies: Additions to the Representations.] Tartu: Eesti Kirjandusmuuseum.

Kallio, Kati & Frog & Sarv, Mari 2017. What to Call the Poetic Form: Kalevala-Meter or Kalevalaic Verse, regivärss, Runosong, the Finnic Tetrameter, Finnic Alliterative Verse, or Something Else? *RMN Newsletter*, No. 12–13, pp. 139–161. Available at http://hdl.handle.net/10138/305420, last accessed on 21 October 2024.

Kallio, Kati & Janicki, Maciej & Mäkelä, Eetu & Saarlo, Liina & Saarinen, Jukka & Sarv, Mari 2023. Eteneminen omalla vastuulla. Lähdekriittinen laskennallinen näkökulma sähköisiin kansanrunoaineistoihin. [Proceed with Care: A Critical Computational Perspective on Digital Folklore Corpora.] *Elore*, Vol. 30, No. 1, pp. 59–90. https://doi.org/10.30666/elore.126008.

Kallio, Kati & Mäkelä, Eetu 2019. Suullisen runon sähköisestä lukemisesta. [On Digital Reading of Oral Poetry.] *Elore*, Vol. 26, No. 2, pp. 26–41. https://doi.org/10.30666/elore.84570.

Kallio, Petri 2015. The Language Contact Situation in Prehistoric Northeastern Europe. In: Robert Mailhammer & Theo Venneman & Birgit Anette Olsen (eds.) *The Linguistic Roots of Europe: Origin and Development of European Languages*. Copenhagen: Museum Tusculanum Press, pp. 77–102. Available at https://www.academia.edu/20252178/, last accessed on 21 October 2024.

Kemppinen, Iivar 1957. *Lunastettava neito: Vertaileva balladitutkimus.* [The Maiden to Be Ransomed: Comparative Ballad Study.] Helsinki: Kirja-mono.

Korhonen, Mikko 1994. The Early History of the Kalevala Metre. In: Anna-Leena Siikala & Sinikka Vakimo (eds.) *Songs Beyond the Kalevala: Transformations of Oral Poetry*. Helsinki: Finnish Literature Society, pp. 75–87.

Krikmann, Arvo 1997. *Sissevaateid folkloori lühivormidesse I: Põhimõisteid, žanrisuhteid, üldprobleeme.* [Insights into the Short Forms of Folklore I: Basic Concepts, Genre Relations, General Problems.] Tartu: Tartu Ülikooli Kirjastus.

Krohn, Kaarle 1903–1910. *Kalevalan runojen historia.* [History of Kalevala Poems.] Helsinki: SKS.

Krohn, Kaarle 1931. *Tunnelmarunojen tutkimuksia 1: Laulusta.* [Studies of Lyrical Songs 1: About Song/Singing.] Helsinki: Suomalaisen Kirjallisuuden Seura.

Kundozerova, Maria 2022. Baza dannykh "Karel'skie runy": Ideia sozdaniia, kontseptsiia, perspektivy. [Database "Karelian Runes": Idea of Creation, Concept, Prospects.] *Al'manakh severoevropeiskikh i baltiiskikh issledovanii*, Vol. 7, pp. 233–240. http://dx.doi.org/10.15393/j103.art.2022.2386.

Kuusi, Matti & Tedre, Ülo 1979. Regivärsilise ja kalevalamõõdulise laulutraditsiooni vahekorrast: Dialoog üle lahe. [About the Relationship of Regilaul-metric and Kalevala-metric Song Tradition: Dialogue over the Gulf.] *Keel ja Kirjandus*, Vol. 2, pp. 70–78. Available at https://www.etera.ee/zoom/22481/view?page=1&p=separate&tool=info&view=0,0,2425,3819, last accessed on 21 October 2024.

Lang, Valter 2016. Early Finnic-Baltic Contacts as Evidenced by Archaeological and Linguistic Data. *Eesti ja soome-ugri keeleteaduse ajakiri / Journal of Estonian and Finno-Ugric Linguistics*, Vol. 7, No. 1, pp. 11–38. http://dx.doi.org/10.12697/jeful.2016.7.1.01.

Lintrop, Aado 2024. Kosmogooniline hari ja selestiline kiik. [Cosmological Comb and Heavenly Swing.] *Keel ja Kirjandus*, No. 3, pp. 219–237. https://doi.org/10.54013/kk795a1.

Mäkelä, Eetu & Kallio, Kati & Janicki, Maciej 2024. Sources and Development of the Kalevala as an Example for the Quantitative Analysis of Literary Editions and Sources. *Digital Humanities in the Nordic and Baltic Countries Publications*, Vol. 6, No. 1, pp. [1–12]. http://dx.doi.org/10.5617/dhnbpub.11517.

Meder, Theo & Himstedt-Vaid, Petra & Meyer, Holger 2023. The ISEBEL Project: Collecting International Narrative Heritage in a Multilingual Search Engine. *Fabula*, Vol. 64, No. 1–2, pp. 107–127. http://dx.doi.org/10.1515/fabula-2023-0006.

Meinecke, Christofer & Wrisley, David Joseph & Jänicke, Stefan 2021. Explaining Semi-Supervised Text Alignment through Visualization. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 28, No. 12, pp. 4797–4809. https://doi.org/10.1109/tvcg.2021.3105899.

Saarinen, Jukka 2006. "SKVR-tietokanta". [SKVR Database.] In: Pasi Klemettinen (ed.) *"Ei se synny synnyttämättä": Selvitys digitointiprojektin vaiheista ja työprosesseista*. Helsinki: Suomalaisen Kirjallisuuden Seura, pp. 36–43. Available at https://www.finlit.fi/wp-content/uploads/2023/10/yht_menneet_elias_loppuraportti.pdf, last accessed on 21 October 2024.

Saarinen, Jukka & Janicki, Maciej & Kallio, Kati forthcoming. Computing Patterned Repetition in Northern Finnic Oral Poetry. *Plotting Poetry*.

Sarmela, Matti (ed.) 1994. *Suomen Perinneatlas: Suomen kansankulttuurin kartasto 2.* [Finnish Folklore Atlas: Atlas of the Finnish Folk Culture 2.] Helsinki: Suomalaisen Kirjallisuuden Seura.

Sarv, Mari (ed.) 2012. *Regilaulu müüdid ja ideoloogiad*. [*Regilaul* - Myths and Ideologies.] Tartu: EKM Teaduskirjastus.

Sarv, Mari & Järv, Risto 2023. Layers of Folkloric Variation: Computational Explorations of Poetic and Narrative Text Corpora. *Folklore: Electronic Journal of Folklore,* Vol. 90, pp. 233–266. https://doi.org/10.7592/FEJF2023.90.sarv_jarv.

Sarv, Mari & Oras, Janika 2020. From Tradition to Data: The Case of Estonian Runosong. *ARV: Nordic Yearbook of Folklore*, Vol. 76, pp. 105–117. Available at https://gustavadolfsakademien.bokorder.se/sv-SE/serie/140/arv, last accessed on 6 November 2024.

Sarv, Mari & Kallio, Kati & Janicki, Maciej 2024. Arvutuslikke vaateid läänemeresoome regilaulude varieeruvusele: "Harja otsimine" ja "Mõõk merest". [Computational Insights into the Variation of Finnic Folk Songs: "Searching for the Comb" and "Sword from the Sea".] *Keel ja Kirjandus*, Vol. 67, No. 3, pp. 238–259. https://doi.org/10.54013/kk795a2.

Seppä, Tiina 2021. Katsooko metsä ihmistä? Suomen Kansan Vanhat Runot ja lajienvälinen kommunikaatio. [Is the Forest Looking at the Human? The Old Poems of Finnish People and the Interspecific Communication.] *Avain: Kirjallisuudentutkimuksen aikakauslehti*, Vol. 17, No. 4, pp. 22–45. https://doi.org/10.30665/av.98306.

Siikala, Anna-Leena 1994. Transformations of the Kalevala Epic. In: Anna-Leena Siikala & Sinikka Vakimo (eds.) *Songs beyond the Kalevala: Transformations of Oral Poetry*. Helsinki: SKS, pp. 15–38.

Sykäri, Venla 2020. Digital Humanities and How to Read the Kalevala as a Thematic Anthology of Oral Poetry. *Arv: Nordic Yearbook of Folklore*, Vol. 76, pp. 29–54. Available at https://gustavadolfsakademien.bokorder.se/sv-SE/serie/140/arv, last accessed on 6 November 2024.

Tangherlini, Timothy R. & Shahsavari, Shadi & Shahbazi, Behnam & Ebrahimzadeh, Ehsan & Roychowdhury, Vwani 2020. An Automated Pipeline for the Discovery of Conspiracy and Conspiracy Theory Narrative Frameworks: Bridgegate, Pizzagate and Storytelling on the Web. *PLOS ONE*, Vol. 15, No. 6, e0233879. https://doi.org/10.1371/journal.pone.0233879.

Tarkka, Lotte 2013. *Songs of the Border People: Genre, Reflexivity, and Performance in Karelian Oral Poetry*. Helsinki: Academia Scientiarum Fennica.

Tedre, Ülo (ed.) 1969–1974. *Eesti rahvalaulud: Antoloogia.* [Anthology of Estonian Folk Songs.] Vols. 1–4. Tallinn: Eesti Raamat. Available at https://www.folklore.ee/laulud/erla/indeks1.html, last accessed on 22 October 2024.

Virtanen, Leea 1987. Suomalaisen ja virolaisen kansanrunouden suhteista. [On the Relationships of Finnish and Estonian Folk Poetry.] In: Leea Virtanen (ed.) *Viron veräjät: Näkökulmia folkloreen*. Helsinki: SKS, pp. 14–34.

Kati Kallio is Academy Research Fellow of the Research Council of Finland at the Finnish Literature Society and holds a Title of Docent in Folklore Studies at the University of Helsinki, Finland. Her research has focused on genres, intertextuality, and performance of Finnic oral poetry in different regions, languages, and historical periods, and on the interaction of oral and literary traditions.

kati.kallio@helsinki.fi

Mari Sarv (Väina) is Leading Research Fellow at the Estonian Folklore Archives of the Estonian Literary Museum, Estonia. Her main field of study has been Estonian and Finnic older singing tradition (runosong), and her focus has been on discovering the modes and layers of variation revealing itself in the abundant material with the help of computational means. Her research also observes the factors impacting folklore archives formation and (re-)use of the materials gathered.

mari@haldjas.folklore.ee

Maciej Janicki is Postdoctoral Researcher at the Department of Digital Humanities, University of Helsinki, Finland. He holds a PhD in computer science with a focus on language technology, and his main current interest is processing of unstructured language data with unsupervised and statistical methods.

maciej.janicki@helsinki.fi

Eetu Mäkelä is Professor of Digital Humanities at the University of Helsinki, and Docent in Computer Science at Aalto University, Finland. At the Helsinki Centre for Digital Humanities, he leads an interdisciplinary research group that seeks to figure out the technological, processual and theoretical underpinnings of successful computational research in the humanities and social sciences.

eetu.makela@helsinki.fi