https://doi.org/10.7592/FEJF2023.90.sarv_jarv

LAYERS OF FOLKLORIC VARIATION: COMPUTATIONAL EXPLORATIONS OF POETIC AND NARRATIVE TEXT CORPORA

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

Risto Järv
Head of the Archives
Estonian Folklore Archives
Estonian Literary Museum, Estonia
risto.jarv@folklore.ee

Abstract: Variation is a core feature of folklore that plays a part in configuring the processes of folkloric communication, transmission and creativity. Computational analysis of large folklore collections offers new outlooks on the study of variation. The article explores the nature of folkloric variation on the basis of folk song and fairy tale text corpora from the Estonian Folklore Archives. Our enquiry into regional variation in Estonian folk song showed different patterns of geographic variation for metre, repertoire and language. To investigate further the possibilities to disassemble the components of variation we turned to a smaller corpus of fairy tale texts from the distinct Seto language community. The results of stylometric analysis of fairy tales collected within the close language community as a rule found that word usage patterns in stories told by the same storytellers were closer to each other than stories with the same content (i.e. tale types). However, other factors, such as collectors' individual styles, recording time, length of text, and storyteller's place of origin contributed notably to similarity as per stylometric scores.

The study has shown that the individual features and aspects of such a complex phenomenon as folklore can follow its own variation patterns. Computational analysis of the variation in large text corpora helps us get a better understanding of the functioning of variation, and the processes of folkloric creation. However, the layers of folkloric variation are not that easy to disassemble, one must be aware of the biases in text corpora and keep in mind the effect of the linguistic variation on the results.

Keywords: fairy tales, folklore, network analysis, runosong, stylometry, variation

INTRODUCTION

Variation is a core feature of folklore denoting diversification of an element or aspect of folkloric expression in its different occurrences whereby these occurrences reveal recognisable similarity. It intertwines the two poles of folklore, creative expression and transmission of existing knowledge and practice. In historical folkloristics, variation has generally been seen as a subtask of classification of folklore texts. However, variation as a phenomenon has rarely caught the specific attention of researchers. Combining large text corpora with the computational analysis might be expected to shed light on the essence and factors of folkloric variation in more general terms. In this article we explore the variation of different aspects of folklore texts in the corpora of Estonian folk songs and fairy tales with the help of various computational methods.

VARIATION AS AN ESSENTIAL FEATURE OF FOLKLORE

Variation is a universal natural phenomenon, and can be seen in the human mind and human expression. Michel de Certeau (1984) claims that the main attribute of cultural transmission is the changing nature of everything that is being passed on. In folkloristics, variation is usually defined as a continuous creative modification within certain limits given by tradition, the concept partly overlapping with improvisation and oral composition (Reichl 2007; Harvilahti 1992; Sykäri 2014). Variation is enacted in the middle field of the main categories of folkloric transmission (stability and innovation, communal and individual, acquisition and expression, knowledge and creativity), and is often shaped by external circumstances such as environment or language. It is a feature that allows folklore to react to events and phenomena with the help of traditional knowledge, to adapt to changing circumstances (Sarv 2008).

Alan Dundes (1989) has claimed that variation is a key concept in folkloristics, a phenomenon that distinguishes folklore from "high culture" and "mass culture". Lauri Honko (2000) states that variation is a defining feature of oral culture, a life-blood of oral tradition, whereas in literary culture, authorship and individual creativity have gained more importance. Juri Lotman (1977) considered the use of variation, instead of unrestricted creativity, to be an aesthetic preference characteristic of folklore, "an aesthetic of sameness" whereby creativity is restricted by given structures, models and rules. Walter J. Ong considers variative expression a practical preference and a tool to secure the preservation of relevant information in conditions of oral memorisation (Ong 1982).

Variation in folklore can be observed both diachronically and synchronically, at individual, situational, communal, regional levels. This complexity can be analysed either through the comparison of selected formalised features of folklore, usually on the basis of recorded performances or by surveying the process of composition and performance of folklore. The study of variation is a multi-layered task of comparison in which various features of folkloric communication (for example content, poetic formulae, meter, language use, melody, performance, functions, communicative aims and modes) and their density dynamics should be taken into account.

In folkloristics, variability has been seen, since the very beginning of the discipline, as an essential attribute of folklore texts; folkloristics has even been termed "the science of variation" (Levin in Beyer & Chesnutt 1997). Despite, or because of, this essentiality, variation as a phenomenon has rarely been the focus of folklore studies (Pöysä 2000).

Research into variation in folklore started in the historic-geographic school in the late 1800s and was mainly used in folk song (Krohn 1926; Kuusi 1949; Dorson 1963; Wolf-Knuts 2000; Tampere 1932; Normann 1935) and folk tale studies (Anderson 1923, 1951, 1956; Uther 2004), developing the concepts of type and variant. The structuralist approach analysed variation for the study of construction principles and the poetic features of folklore phenomena (Bogatyrev & Jakobson 1972; for example Propp 1968 [1928]; in folk songs, for example Steinitz 1934; Anderson 1935; Sadeniemi 1951; Leino 1970; Sarv 2000).

One of the most serious theoretical attempts to approach folklore variation as a phenomenon was Walter Anderson's Law of Self-Correction and his experiments on the transmission of folklore (Anderson 1923, 1951, 1956; see also Seljamaa 2005; Hafstein 2001). Oral-formulaic theory, as developed by Milman Parry and Albert Lord, saw variation as a natural result of oral composition (Parry 1930; Lord 1960; Foley 1985; Ong 1982; on application to Estonian folk songs, see Kolk 1962; Tedre 1964; Harvilahti 1992, 2004).

By the 1970s, the focus in international folkloristics shifted to communication, context and creative production. Variation was reconsidered as a tool of adapting folkloric knowledge according to a particular situation, presenting performers with the opportunity to express their creativity (Bauman 1984; Hymes 1981; Tedlock 1983; Foley 1992, 1995; Kaivola-Bregenhøj 2000).

For research using the methodology of the historic-geographic school, the existence of large text collections was critical in order to figure out, on the basis of comparison of variants, the historical spread and developments of plots and their hypothetical archetype or 'original form'. This aim led collectors to focus on texts and to carefully record even slightly differing versions of plot and wording. In the later periods, folkloristics has turned more to methods focusing on individuals and performance, and previous text-centred collecting principles have come under severe criticism. The potential of large historical text collections has remained untapped by and large. The interest in formulaic language and composition seems to have got stuck in the missing ability to process large amounts of data.

The introduction of computers to humanities research created new perspectives in the study of large folklore text corpora. Arvo Krikmann has shown that, similarly to language (and other natural phenomena), Zipf's law applies to archival collections, for example there are a very few very popular proverbs, and very many proverbs that have been recorded only once or twice. Krikmann explained this regularity in distribution in folklore collections (and in live communication) in terms of the transmission of knowledge and creativity: widely known texts form the core of the genre, are more stable, safely transmitted, and function as a model for new texts, while the peripheries are characerised by live creativity and testing the acceptance of texts (Krikmann 1997). Materials in the large archives are able to elucidate variation as a process and give us hints as to how and why "types" (or groups of similar items) emerge in folkloric communication (Hiiemäe & Krikmann 1992).

With the development of computational methods, especially in the field of natural language processing, data mining and the use of geo-information systems, as well as the ever broadening digital availability of source materials, the field of digital folkloristics has grown alongside the general flood of digital humanities (e.g. Abello et al. 2012). There is a rich variety of possibilities for the application of digital methods to large archival corpora in order to advance folkloristic research.

Although the field of folklore research has increasingly become the examination of the relationship between the individual and his/her folkloric expression, it is important to know the mechanisms of variation in order to understand the essence of folklore. Computational models based on large material collections can help us better understand the patterns of formation of collective thinking and memory (see for example Tangherlini et al. 2020).

In the following we will observe the different layers of variation in our research material, Estonian folk songs and fairy tales, in order to find out to what extent the dynamics of similarities and differences in a text collection reflects variation in language, style, ways of expression, and content. The methodological challenge in studying the Estonian folklore text corpora derives from the highly variable non-standard language of folklore records, which tools designed for the standard language can neither automatically lemmatise nor grammatically analyse. Moreover, compared for example to English, Estonian is morphologically much more complex: words usually have a number of morphological forms, sometimes along with stem variation. Dialectal variation involves all the levels of language. Estonian has two main dialects, South Estonian and North Estonian, both with several subdialects. In computational analyses we always have to bear in mind that linguistic variation contributes to a considerable extent to text variation.

REGIONAL VARIATION IN ESTONIAN FOLK SONGS

The current chapter focuses on an archaic folk song tradition called runosong[1] that has been shared by several Finnic peoples, and is characterised by a specific poetic structure that combines a specific meter with a trochaic core, alliteration and parallelism as key features. Runosong has been considered a way of expression that characterises Finnic peoples, and therefore the songs have been transcribed in large volumes, with the peak at the end of 19th century, and stored in the archives in Estonia, Finland and Karelia. The singing tradition has, in the majority, faded away along with the progress of modernity, but the archival collections have been in lively re-use by composers, musicians, writers, etc. Runosong has been at the focus of folkloristic research in the respective countries, and to date most of the texts have been digitised and brought together in well-organised databases that are available for computational research (ERAB 2023; SKVR 2021; cf. Järv 2016: 33–34; Sarv & Oras 2020). Linguistic variation, however, poses a noteworthy challenge to this in that the corpus is multilingual containing songs in several Finnic languages and their dialects. Runosongs use a specific idiom that in some regions differs from the spoken language in several respects. Recently, the FILTER project, funded by the Finnish Academy, has made significant steps in dealing with this question (see Janicki et al. 2022; Janicki 2022).

The issue of layered variation caught our attention in connection with the study of metric variation in Estonian runosong (Sarv 2008, 2015, 2019), which revealed clear regional variation (Figure 1). One might expect that metric regions would reflect the tradition areas of runosong in more general terms, but the pattern diverges from the general idea of Estonian tradition regions (Figure 2) or dialect areas (Figure 3) with the metric regions crossing the main dividing line between South and North Estonian linguistic and cultural areas. This prompted us to observe patterns of regional variation in two other aspects of runosong: (1) typological distribution, and (2) distribution of most frequent word forms (stylometric analysis).

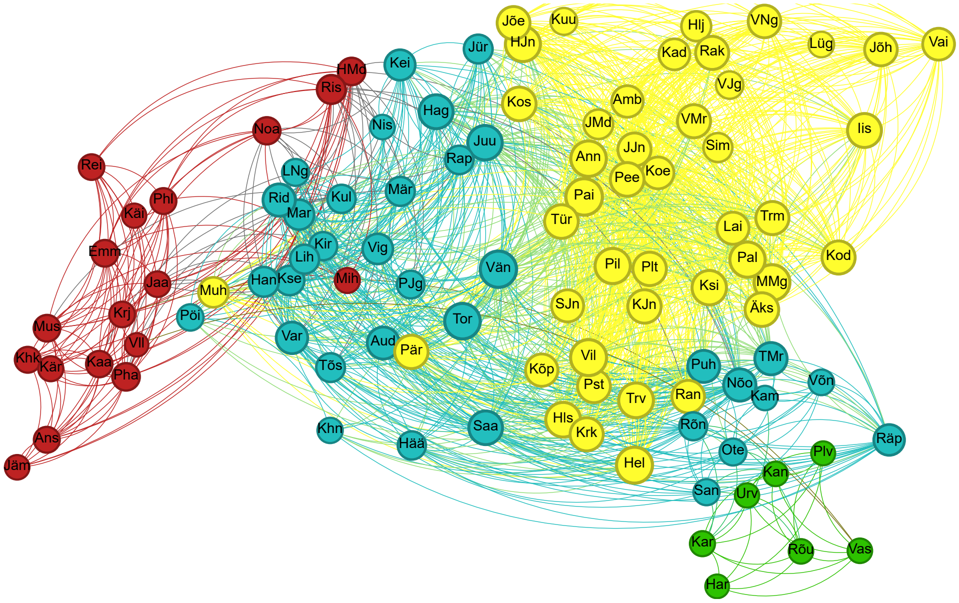

Figure 1. *Metric areas of Estonian runosong. Circles show parishes grouped together using the network modularity algorithm (Blondel et al. 2008) used in Gephi application (Bastian et al. 2009). The metric proximity of the parishes is calculated by summing the differences of percentages of lines with 7 different metric features (Sarv 2008, 2015) between each pair of parishes (map from Sarv 2019: 108).*

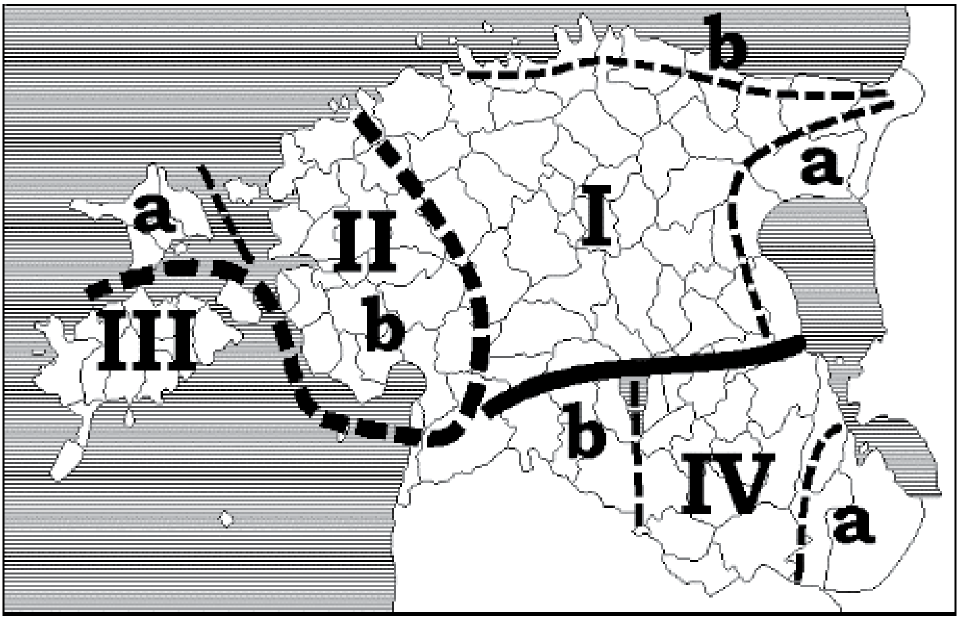

Figure 2. *Estonian ethnographic and folklore regions according to Oskar Loorits: I north Estonia, II west Estonia, III Saaremaa island, IV south Estonia (map from Krikmann 1997).*

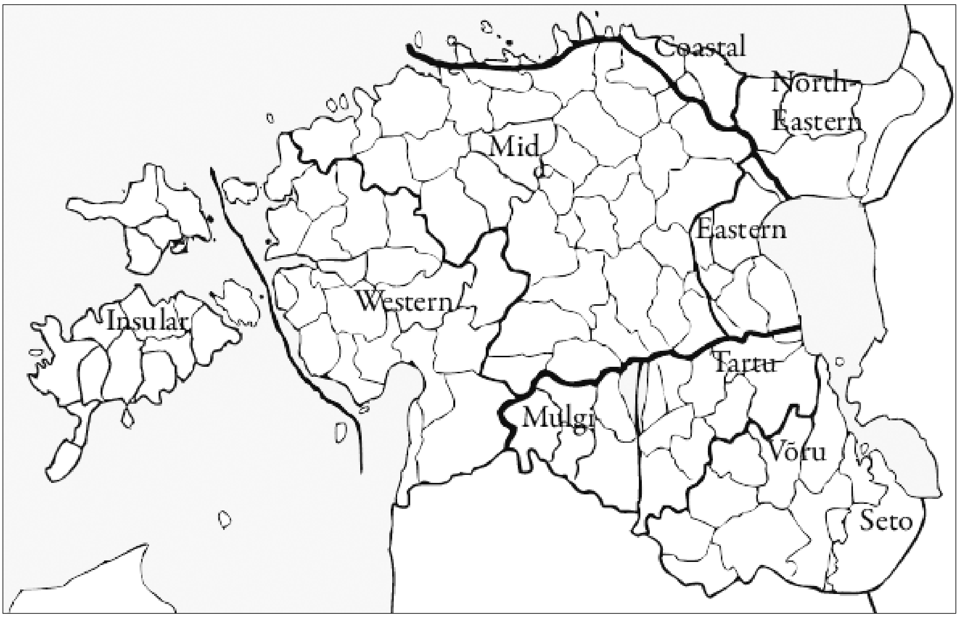

Figure 3. *Map of Estonian dialects:*
1) North Estonian dialect group: Mid, Eastern, Western, Insular dialects;
2) South Estonian dialect group: Võru, Mulgi, Tartu, Setu dialects;
3) North-Eastern Coastal dialect group: North-Eastern, Coastal dialects (Lindström & Pajusalu 2003: 242).

In order to analyse the regional distribution of song types we used the database of Finnic runosong compiled within the framework of the FILTER project, specifically the Estonian part deriving from the Estonian runosong database. The FILTER database as a whole currently contains around 100,000 Estonian texts, the great majority of which are runosongs, although it also includes a significant number of texts from other genres. Every text is usually localised to one or two parishes. For the analysis of typological distribution we used the data on folkloristic song types. Originally the type assignments were digitised along with texts from the machine-typed copies of manuscripts at the Estonian Folklore Archives, which were scanned and OCR interpreted for the database (cf. Järv & Sarv 2014; Sarv & Oras 2020). The mechanical transcription of songs took place over the ca 70 years of the folklore archive's history, and the classification was not consistent. Since digitisation the classification in the database has been revised, although this is still a work in progress: currently approximately half of the Estonian texts in the database have a revised type assignment, with all together 1943 different song types assigned. We considered this number to be sufficient to uncover how Estonia divides into regions according to song reportoire.

In order to identify regional divisions we first calculated the similarity scores for each pair of the 105 parishes in Estonia (the parish island of Vormsi was excluded as the runosong tradition was unknown among the Swedish population of the island). We followed the methodology applied by Arvo Krikmann in various folkloristic and linguistic geo-labelled data collections (for example Krikmann 1980, 1997, 2014). The procedure for measuring the similarity between two regions is based on a comparison of the share of their common repertoire in the total corpus with the shares of the repertoires of both regions in the total corpus. In other words, the real size of the intersection is compared with the expected size, taking into account the total folkloric capacity of the regions in question. The song types are of very different popularity, and thus also frequency in our collections, and it is not easy to find a procedure to balance an estimation of very rare and very frequent types.[2] As a basis of regional division we chose to observe only the geographic spread of each song type, not the number of texts collected.

For each pair of parishes we found the similarity index *sim* by subtracting the expected probable number of common song types from the actual number of common song types.

$$sim = \mathrm{CT}–\mathrm{ECT}$$

The expected probable number of common song types for two parishes ECT is calculated by multiplying (1) the proportion of song types occurring in parish 1 among all the combinations of type and parish by (2) the proportion of song types occurring in parish 2 and by (3) the number of all the different combinations of type and parish.

$$ECT = \frac{number\ of\ types\ in\ parish\ 1}{\sum_{n=1}^{105} number\ of\ types\ in\ parish\ n} \times \frac{number\ of\ types\ in\ parish\ 2}{\sum_{n=1}^{105} number\ of\ types\ in\ parish\ n} \times \sum_{n=1}^{105} number\ of\ types\ in\ parish\ n$$

The parish pairs with a positive *sim* have more common song types than would have been expected from their total size of repertoire, while the parish pairs with a negative *sim* index had fewer common types than expected.

As some of the song types are widely known, most of the parish pairs had mutual connections that formed a dense network. For network analysis we used the Gephi application (Bastian et al. 2009), and for community detection the modularity analysis method implemented in Gephi (Blondel et al. 2008).

The parishes on the graph were geo-located using the Gephi GeoLayout plugin by A. Jacomy. For the poetic meter of runosong, network analysis is proven to give more clear-cut results in detecting regional division (Figure 1), compared to other methods applied (see Sarv 2008: 44–45).

As a result of the analysis, Estonia was divided into three relatively coherent areas that differ considerably by song repertoire: (1) west Estonia, (2) south Estonia, and (3) north-east Estonia (Figure 4). The border of the south Estonian area almost overlaps the main dialect border that separates the South Estonian and North Estonian cultural and linguistic areas. The border between western and north-east typological regions does not overlap the dialect border, although it does overlap the main metric division between the western and eastern regions. The geographic coherence of the network communities obtained as a result proves the relevance and suitability of the chosen data analysis methodology.

The extent of variation also depends on collection density: in regions with a richer tradition during the period of active folklore collecting, the nature of the material is more variable; in regions where songs were collected during the phase of fading tradition, only the most popular and custom-bound songs survived long enough to be collected. The geographic outliers tend to be anomalous in terms of collection density.

In general the results are congruent with Krikmann's generalisation of the geographical distribution of Estonian dialect words, proverbs and riddles, as based on archival collections. The material that Krikmann used gave three main dialect and tradition areas:

1) The South Estonian dialect area;
2) The western islands and a large part of western Estonia;
3) The rest of the North Estonian dialect area.

(Krikmann 2014: 63).

The network picture, configured by similarity measures (Figure 5),[3] shows that western, northern, and even southern areas seem to be connected via the south-western county of Pärnumaa. We can hypothesise that this might have a natural background in the Pärnu river, which functioned as a communication route from the south-west towards the north-east in pre-modern times. It is logical to assume that ease of connection between regions would result in a common song repertoire, and thus also in the closeness of these parishes in the network graph. In addition, there seems to be an anomalously large divide between western and northern groups. We can speculate that the reason for this could be the location of the capital city Tallinn on the northern shore, between the two regions, which since the Christian invasion in the 13th century had mainly been inhabited by foreigners (Germans, Russians) and probably did not foster the spread of folklore.[4]

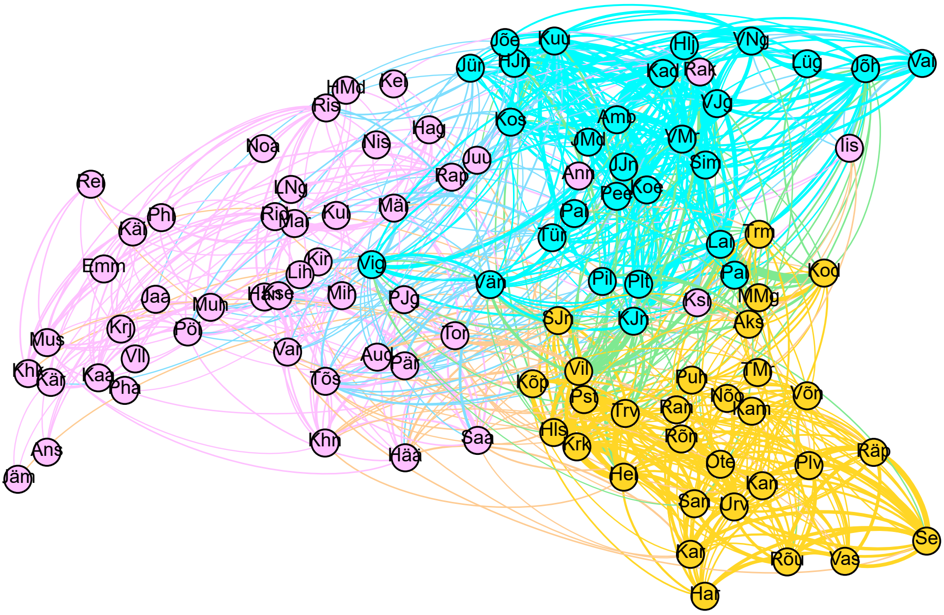

Figure 4. *Regional division in the Estonian runosong area on the basis of typology data.*

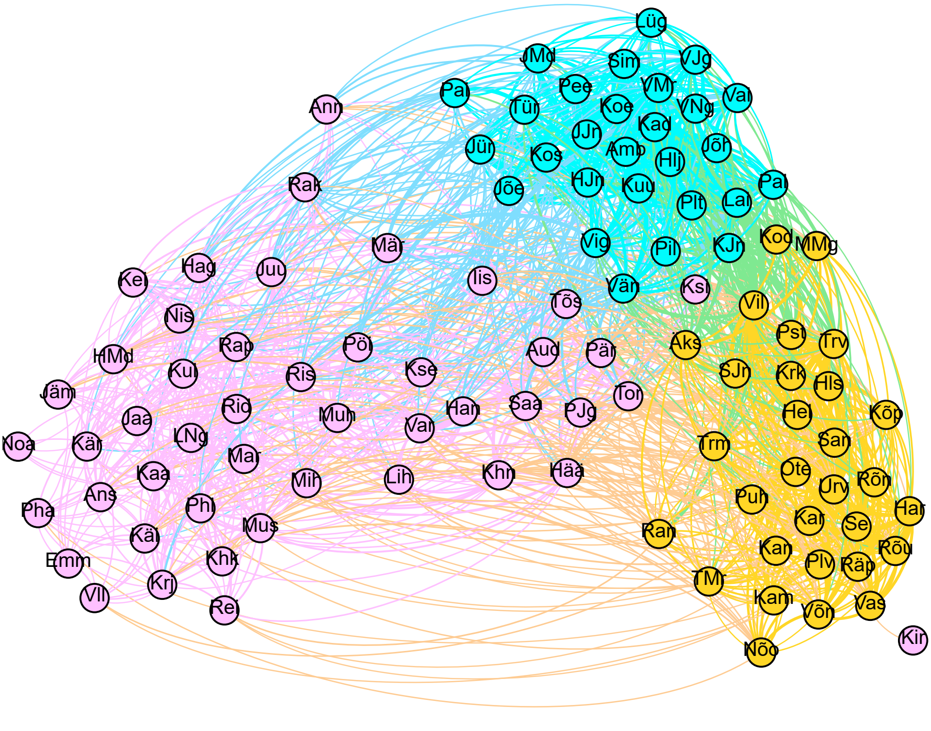

Figure 5. *Network of Estonian parishes on the basis of typological similarity measures. The parishes of Pärnumaa county, located in the graph in the middle of all the three groups, are circled in red.*

For the distribution analysis of the most frequent word forms we used stylometry and R package 'Stylo'. Stylometry is a method of assessing similarities between texts on the basis of the use of frequent words or other units (such as characters, combinations of two or more characters, combinations of two or more words, parts of speech tags, or other information). Stylometry is widely used for authorship attribution and in explorations of the stylistic variation of text collections (on the method as well as the Stylo package, see Eder et al. 2016). The method has also been used for the analysis of language change (Eder & Górski 2016) and variation (Mäkinen 2020). In our case, we expected stylometric analysis first and foremost to reflect dialectal variation rather than stylistics. Estonian can be divided into three main dialects, South and North Estonian, and the North-Eastern Coastal dialect, each of which is divided into several sub-dialects. Runosong language forms a specific language register, using systematically archaic word forms to conform to the requirements of meter. The major changes in the prosody and syllabic structure of Estonian that took place around 500 years ago are only partly reflected in runosong language, depending on the region (Sarv 2019). The aim of our observation was to detect runosong language regions with the help of stylometric analysis.

Proceeding from our aim for regional analysis, we decided to compile analysis files for every parish consisting of all the texts from each parish in the database. Thus, we organised the Estonian runosong texts labelled in the database with a parish name into a corpus consisting of 105 parish files with an average length of 47,880 tokens (ranging from 560 to 192,678). The stylometric analysis of the 1,000 most frequent word forms was performed by Stylo using the classic delta measure. The similarity measure sums the differences of relative frequencies for each word in both texts, normalised by standard deviation of relative differences in the same word in the corpus (Burrows 2002; Šeļa 2021). The network edges output was processed using Gephi, and the same procedures as in the previous case. The results are depicted in figure 6 (a network drawn on the basis of similarity measures) and figure 7 (a geolocational graph of the same data).

The stylometric analysis reflects a clear division between South Estonian and the northern dialects (North Estonian and the North-Eastern Coastal dialect), with relevant subgroups in both areas. Figure 7 shows that the groups in the network are also geographically coherent, thus the analysis reflects geographic (and presumably linguistic) proximity. In the case of South Estonian, one of the four main sub-dialects (Mulgi) is comprised of a separate subgroup. In the case of northern group, the network division very roughly overlaps the dialect division, i.e. the North-Eastern Coastal dialect, the Mid dialect, the Western dialect, and the Insular dialect areas form separate subgroups within the network. However, there is a separate subgroup in the border area with South

Estonian which in dialect division is split between the Eastern, Mid and Western dialects. Here the stylometric analysis shows probable influence from the frequent use of refrain words in this area. Songs with refrains are known in the South Estonian dialect area as well as in those parts of the North Estonian border area that are next to South Estonia, usually appearing in songs related to either calendar or family customs. Refrain words can occur among the most frequent words, as they are repeated often.

To find which word forms are peculiar to each region, we ran a keyness analysis for all the groups[5] and observed 20 positive keywords for each group. We expected the keywords generally to be dialect variants of grammatical words as they usually dominate among the most frequent words in any text. About the half of the keywords in each regional group were grammatical words, for example various forms of the verb 'to be' (*on*, *one*, *o*, *oo*, *uo*, *om*, *pole*, *põle*, *õli*, *olid*, *oll*, *oleks*, *ollin*), and more occasionally 'this/that' (*sie*, *see*, *sii*, *tuu*), 'on' (*peal*, *peale*, *piäle*, *pealla*, *peele*, *pääle*), 'there' (*seal*, *siel*), 'me' (*mina*, *minu*, *mind*, *moole*), 'we' (*me*, *mi*, *mii*, *mede*), 'then' (*siis*, *sis*), 'no' (*ei*, *es*). In addition, among the keywords were grammatical words that are used as filler particles in regional performance traditions: *iks/õks* 'ever', *no* 'now' in the south-eastern group, where this phenomenon mainly relates to the special characteristics of the Seto singing tradition with longer melodies and abundant filler words; *ja* 'and', *aga* 'but' in the western and insular groups, where the usage of these small words relates to a newer, rhythmically more complex performance style (Rüütel 2012).

Keywords, however, also reflect content words that are peculiar to each region. 'Mother' is one of the most central concepts in Estonian runosong, used along with rich poetic ornamentation. Keywords contain the regional dominant variants of mother (*eit*, *eide*, *eidekene*, *emm*, *emä*, *emakene*, *memm*, *memme*, *memmekene*, *ennekene*, *ime*, *imä*). The words *tere* 'hello' and *aitimal* 'thank you' among the keywords in the insular group illustrate the use of songs in live communication as runosong has traditionally been part of ceremonies, most notably weddings and mumming processions. The songs from the western group and Mulgi group contain the other keywords related to these customs (*langud* 'in-laws', *mart*, *marti* or *märti* 'mummers'). The central northern geographical area is known for its developed swinging tradition, and this is reflected in keywords from this area: swing (*kiike*, *kiige*) and swing smiths (*kiigesepad*). As we supposed, in the border area between the North and South dialects there are several refrain words among the keywords (*kaasike*, *kaske*, *kaanike*, *kaine*, *nuku*, *lõpele*).

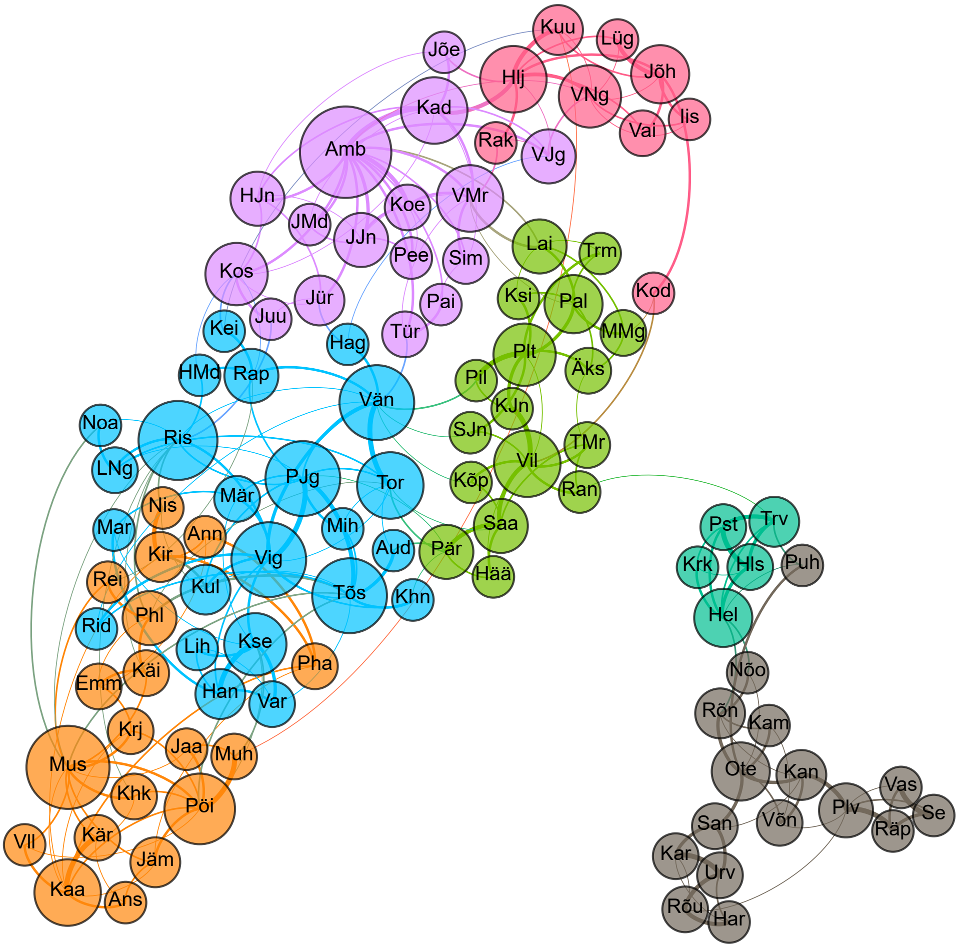

***Figure 6**. The results of stylometric analysis of the 1,000 most frequent word forms in the runosong texts that have a parish label in the Estonian runosong database, depicted as a network based on the similarity measure classic delta and grouped according to modularity analysis, as implemented in the Gephi program.*

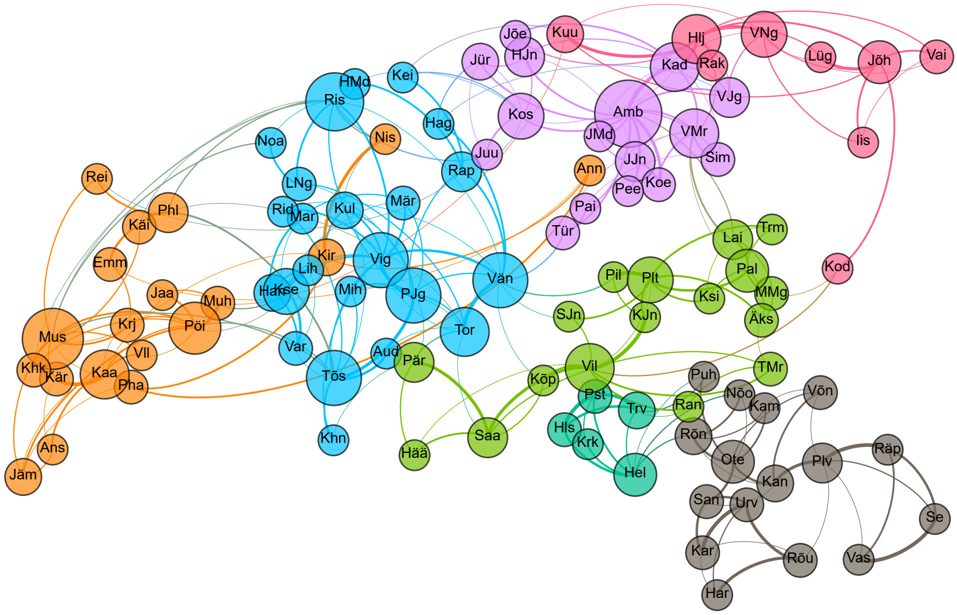

Figure 7. *The results from Figure 6, arranged according to parish location.*

According to these keyness observations we can conclude that stylometric analysis of the use of word forms, and the regional groups obtained, in the first instance reflect dialectal variation, but to a notable extent also derive from other aspects such as the specific features of performance tradition like filler words and refrains, communicative strategies (interaction vs narration), and the dominant genre of the region.

The study demonstrated that the regional peculiarities of runosong have formed as a combination of various features: the musical structures of songs, metric peculiarities, dominant genres, customs and performance habits along with communicative modes related to them. The observations showed runosong variation as a multi-layered phenomenon with the analysis resulting in geographically coherent regions in all the three observations, although in each case the outcome was different. It is not a trivial task to discover the effects of individual features that do not always follow the same regional patterns in the conditions of severe linguistic variation. Linguistic variation always underlies and contributes to textual variation in folklore, sometimes also motivating components of poetic structure.

STYLOMETRIC ANALYSIS OF SETO FAIRY TALES

As we saw from the previous part of our analysis, variation in folklore unfolds as a complex phenomenon, with different components having different variation patterns. It is often hard to discern these components in analysis. In order to be able to distinguish better the individual layers of variation, we decided to investigate – using texts from the same region and language group – if, and how much, the proximity measures gained as a result of stylometric analysis reflect the content of the texts, or of the performers' individual styles (as would be expected in case of stylometric analysis).

Runosong texts in general tend to be short, the average length in the Estonian part of the FILTER database currently being 68 words (stylometry is usually said to give more reliable results for longer texts).[6] Therefore we chose longer narrative texts for further analysis.

The database of Estonian folktales (cf. Järv 2016: 38–40) currently consists of 16,000 texts along with metadata, including approximately 6,000 fairy tale texts (ATU 300–749). About one quarter of the material comes from the Seto region at the south-eastern border of Estonia with its specific culture and language variant, a sub-dialect of South Estonian. An overview of the Seto fairy tale tradition can be found in *Setu lauludega muinasjutud* (Salve & Sarv 1987: 10–16, English summary 198–199), an academic publication on Seto fairy tales with songs.

The tradition of telling fairy tales has been preserved in the Seto region for longer than elsewhere in Estonia, which is also the reason for the abundance of Seto material in the archives. On average, the texts recorded in the Seto region are also longer and more elaborate. We chose for our analysis 125 texts with an average length of 1,409 words. Most texts (109) in this selection come from the folklore collection of Samuel Sommer, who also aimed to collect the life stories of the tradition bearers. This plan was fulfilled only partly (see Kalkun 2011: 193–195, 199–201), although luckily enough at least the name, age, birth and residence of the storytellers was noted down along with the texts, which does not apply to all fairy tale texts in the collections of the Estonian Folklore Archives. These texts were recorded by volunteer collectors, as are four texts from the same period from the collection of the Estonian Folklore Archives (ERA). Twelve texts, approximately one tenth of our sample, were collected by professional folklorists between 1946 and 1953. The number of texts in our selection is not large, but they come from the same linguistic area and are provided with basic metadata.

Classification of folklore texts according to their content has for a long time been one of the basic methods of folkloristics. The best known are international folk tale typologies, first compiled by Antti Aarne (1910) and enlarged by Stith Thompson (1928, 2nd revision 1961), and most lately by Hans-Jörg Uther (2004), although the method has been used to classify other genres of folklore as well, such as song, legend, proverb, riddle. Despite criticism that the classification is based more on characters than plots (Dundes 1989) the system is used widely among the scholars of folk tales to this day. At the same time, it has been also a cornerstone for the historic-geographic method, which aimed to find the original form and home of each plot, as well as later additions. Although the aims of the historic-geographic method have generally fallen into disuse, the typological classification has retained its function as a useful tool in gaining an overview of the large number of records. It enables us to examine the geographical spread of a folkloric type and observe its variability, which in turn reveals characteristics of folkloric communication. When observing the voluminous folklore collections and archives it is evident that folkloric types are not a construct of the researchers, but rather are a reality typical to folkloric communication as texts (and melodies) tend to group into sets with the same or similar form and/or content. As typological classification in itself is a method to cope with big data, it would require a reasonable number of texts to find what is typical and what is exceptional. Today typological classifications can be used as source data for computational analysis, combined with other dimensions of folklore texts, for example their geographical spread, their linguistic properties, their performers.

For the current analysis we chose 125 Seto fairy tale texts from the database according to the following principles:

1. The texts were transcribed from the 22 storytellers who have the largest number of fairy tales in the manuscript collections. As a rule these people have also been versatile tradition specialists. In top spot are Maria Kütte, from whom the collectors recorded 187 fairy tales in total as well as a large number of songs (SNE 2014: 33), and Feodor Vanahunt with 176 stories (Kalkun 2015: 9). For most people included, the number of the recorded fairy tales is considerably smaller (~20).
2. Among the stories from these storytellers only fairy tale types that have been popular in the Seto region, and have been transcribed from several storytellers, have been included in our research corpus[7]. Most of these tale types are known more widely in Europe while some are found only in the Seto region. They may have been present more widely in Estonia as well, but the only versions to make it into the archives are from the Seto region. Compared to the animal tales (which are also popular in the Seto region), Seto fairy tales are longer and thus better suited to our study. Each type in our selection has at least 3 different texts; types ATU 480, 572*, 613 and 700 have more than 10 texts.

3. In our sample corpus we selected only the texts that consist of one tale type. Although types are often mixed and combined in the tradition as well as in the archival records, this decision was made with the intention of having a clearer understanding of the effect of the content of the tales on the results.

The sample consists, thus, of the 16 most popular Seto fairy tale types from the 22 storytellers who have the most fairy tale records in the archives.

For the analysis, all the information outside of the main text, such as titles, comments, final comments, was removed as it is not always clear if it was in the original story or added by the collector. The effect of the collector's linguistic style and ability, and the steps that form his or her method of writing down the text, upon the collected texts, has been discussed previously, especially in the case of the Sommer collection, where the collectors were paid and were thus motivated to produce more texts (Kalkun 2011: 195–199). We labelled every story with the surname of the storyteller and a four-digit ATU type number, for example Huntsaar_0300 is the variant of The Dragon Slayer (ATU 300) told by Vassä (Vassilissa) Huntsaar. If there were several variants of a story told by the same person, we numbered the variants with roman numerals (for example, 0480_I, 0480_II).

Stylometric analysis of the 100 most frequent word forms was performed using Stylo's classic delta measure. Figure 8 shows a network graph using Gephi, with nodes placed on the basis of similarity measures at network edges output of Stylo as in the case of previous runosong analysis. Placement of the texts in the graph reflects similarity in use of the most frequent word forms. The graph is coloured according to the Gephi modularity analysis, which divides the network into more densely connected parts. At the edges of the graph we see six more distinct clusters (with more distinct language use) consisting of tales from one or two storytellers. In the centre the grouping is less clear, but the tendency still seems to be the same – the detected groups generally formed from the tales of a small number of storytellers. In stylometric analysis the usage pattern of most frequent words is considered to be a characteristic feature of an author's individual style and thus we can expect, in the case of fairy tales, stories by the same author to cluster together.

However, if we look at the same graph coloured by storyteller (Figure 9), we see that the picture is not so even. We can see, especially in the central area of the graph, that there are quite a number of outliers, i.e. stories told by the same storyteller that are placed further away from the main body of his or her stories. In addition, the stories by storyteller Maria Kütte (together with the stories by Maria Laanetalu) form two close but still distinct communities in the network.

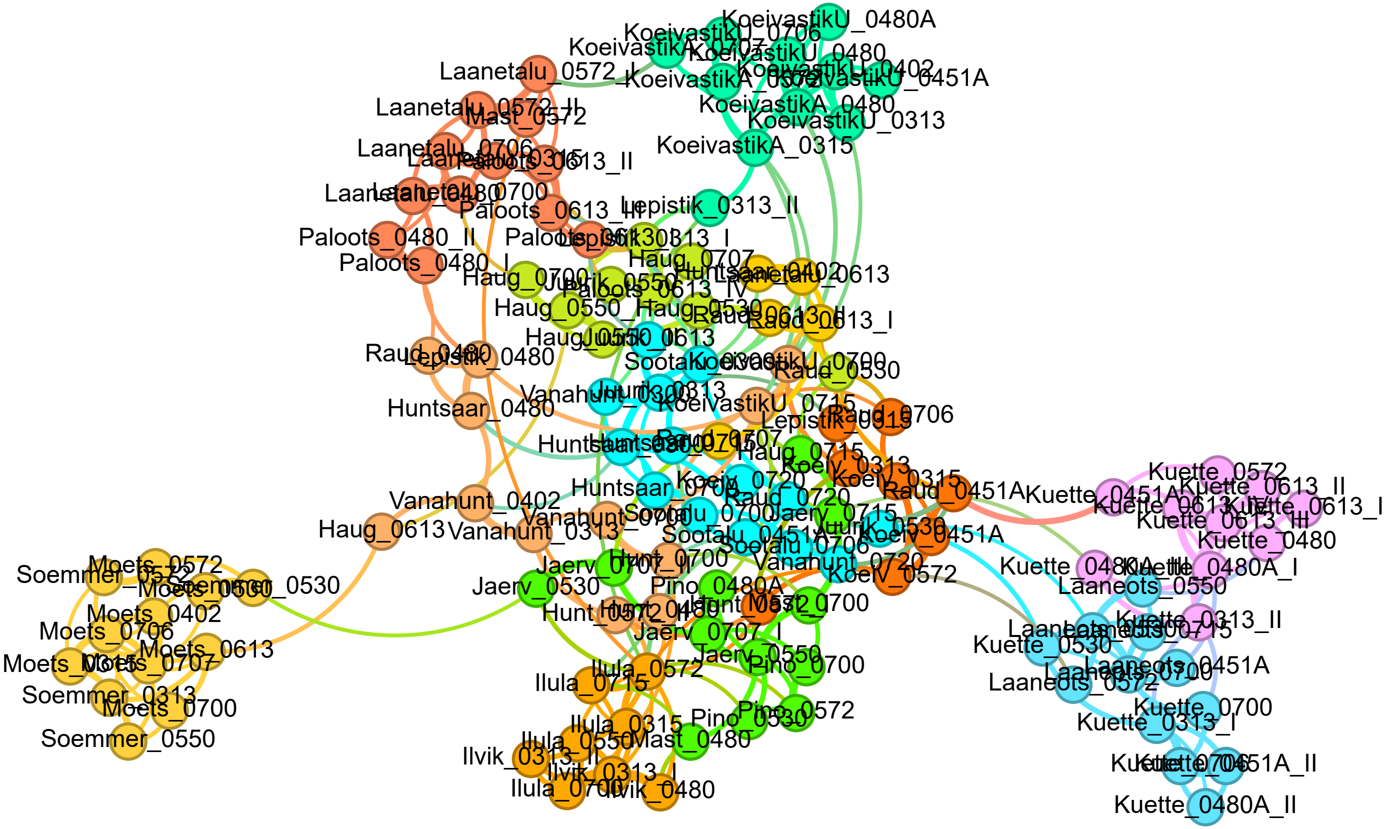

Figure 8. *Fairy tale network based on similarity measures from Stylo network output. The colours represent groups from the Gephi modularity analysis. The labels combine name of storyteller and tale type number.*

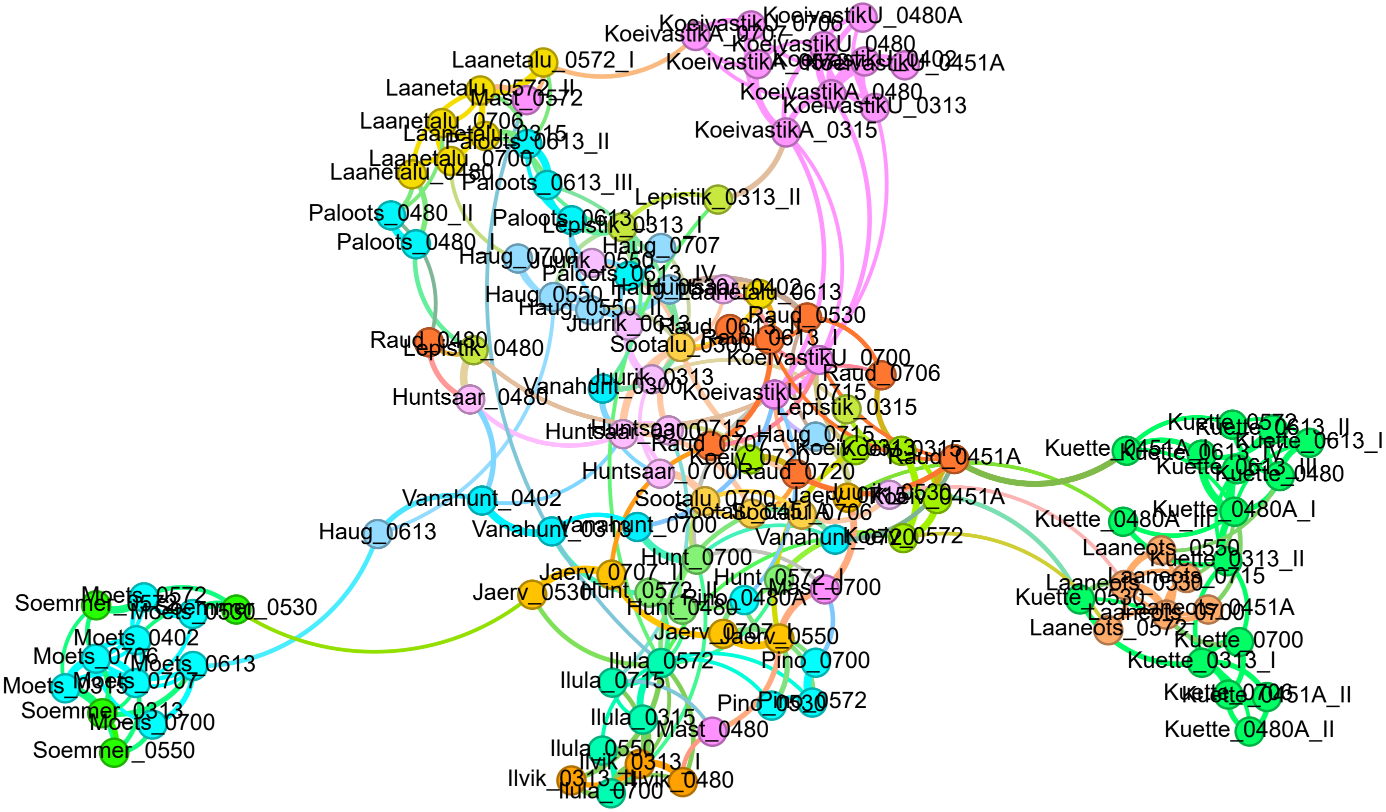

Figure 9. *Network of fairy tales with storytellers identified by colour.*

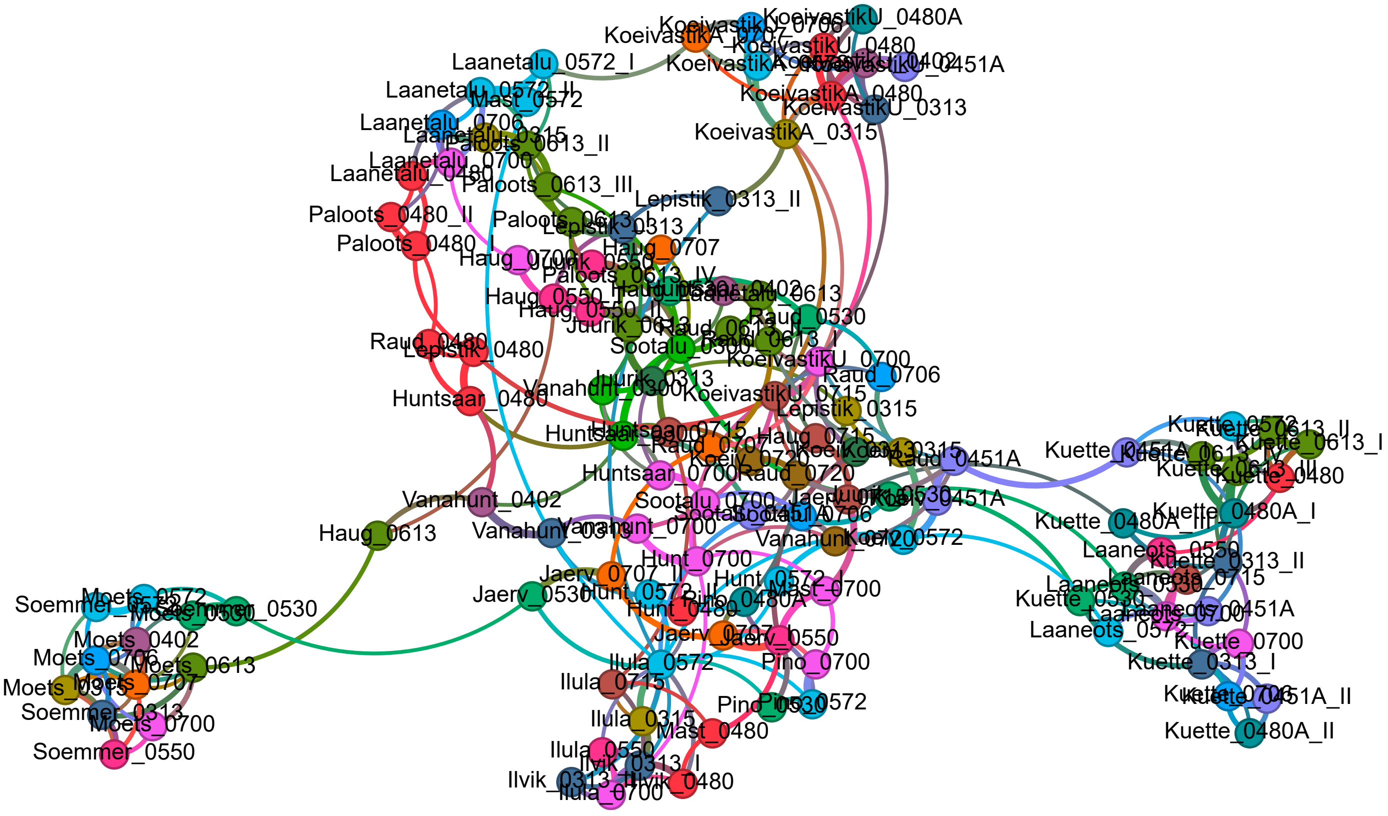

***Figure 10.** Network of fairy tales, tale type identified by colour.*

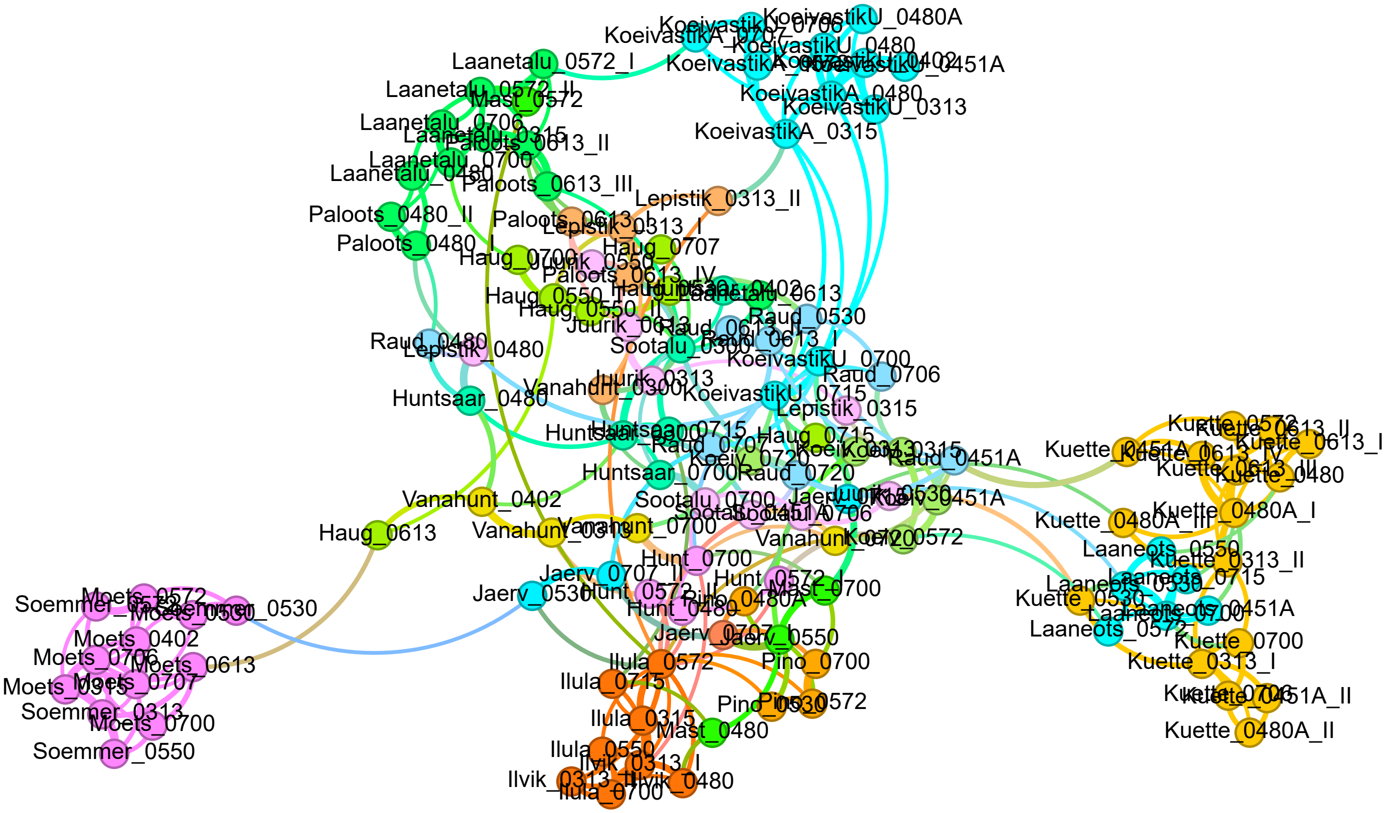

Figure 11. *Network of fairy tales, collectors identified by colour.*

We wanted to find what factors affect similarity in the use of the most frequent word forms, why stories from some storytellers form a distinct group and the others do not, what brings the stories of two storytellers together in a group at the edge of the graph, why the stories of one storyteller form two distinct groups, why some of the stories lie outside other stories by the same storyteller. Using the available metadata we could further check the effect of content, collector, linguistic variation, collection year and length of the story.

In figure 10, the graph is coloured by tale type (ATU index references) with each of the 16 types having its own colour. From the placement of the coloured circles we can see that in general the tale type is definitely not the cause of similar use of the most frequent word forms, and thus, as a rule, such an analysis does not bring together stories with similar content (except for cases of a storyteller telling one story several times, for example four variants of ATU 613 told by Maria Kütte, marked in dark green, cluster together). However, on closer inspection there are quite a number of instances when stories of the same type told by different storytellers occur next to each other, most notably a group of five variants of ATU 480 in the upper left corner (marked in red), as well as for example the same type told by Mast and Ilvik at the bottom of the graph, or type 451A by Raud and Kõiv (lilac) to the right of the central cluster, etc. Closer observation of ATU 480 variants, the Story of the Good and Bad Girl, reveals that this could be related to the frequent use of the *Kulla tütrik, tsirgu tütrik* ('dear girl', 'lovely girl') address formula, which multiplies the frequency of these words. Thus we cannot say that the content of the stories plays no role in similarity measures. Formulaic language characteristic to folklore has a particular impact on the word frequency structure.

As stories were recorded in writing by collectors from live performances, rather than by storytellers themselves, we can expect that in many cases these stories have been not written down word by word. It was usual to make more or less complete notes in the field, and write down a polished version of the story afterwards. In any case the collectors have contributed to the final version of the tales in the archives, and this may have an effect on the style and wording.

Figure 11 identifies stories by collector. In our sample the tales were collected by 19 different collectors, some of whom collected stories from only one storyteller while others from two people; one collector, Aleks Põhi, transcribed stories from three different performers (marked in pink in the upper part of the graph). We see in figure 11 that most of the distinct clusters at the edges represent the collections of individual collectors: the light blue group of stories were recorded in writing by Theodor Kõivastik from the same storyteller Anna Kõivastik; the green group was recorded in writing by Viktor Ruusamägi from Maria Laanetalu and Anastasia Paloots (which are in proximity to other stories in other groups by her recorded by Põhi); the lilac group was recorded in writing by Mihhail Pihlapuu from Eudokia Mõts and M. Sõmmer; and the orange group are stories from Evdakia Ilula and Vassili Ilvik, recorded in writing by Paul Külaniit. The latter two collector groups also overlap with the Gephi network clusters. In the middle of the graph we can also see a tendency of stories recorded by the same collectors to cluster together. It is also noteworthy that 11 stories (all except one) collected by professional folklorists Selma Lätt, Veera Pino and Herbert Tampere in the 1940s and 1950s from three different storytellers are close according to their word use and gather in the same network cluster (green in figure 8). It is possible that the writing style of pre-war period amateur collectors and professional folklorists of the post-war period have some systematic differences. The proximity of word use in stories by Maria Kütte and Maria Laaneots (on the right side of the graph), as recorded by two distinct collectors, still cannot be explained.

It was evident from the study of runosong that the texts reveal clear patterns of regional variation. We aimed to exclude this by selecting fairy tales from one language region (although the Seto language also has its own dialects). We do not have any data on the dialect or subdialect of the fairy tales. However, the recordings have information on the storytellers' origin locations, according to which the graph is coloured in figure 12. Most of the storytellers in our sample come from the Vilo community (yellow in the figure). There are considerably fewer stories by storytellers from Mäe, Meremäe, and Saatse communities, although the stories from the same region, in general, tend to cluster together. The placement of stories from these four communities on the graph have a certain geographical logic. Mäe and Vilo communities are located furthest from each other, Saatse and Meremäe are in between. Among the overall yellow area of the network, the few outliers, stories that cluster together by type (independent of storyteller, collector or location of origin) are now clearly visible. We note that the storytellers of the rightmost distinct group both come from the Mäe community, which is the centre of the northern Seto dialect (Hagu & Pajusalu 2021).[8] Thus, it seems probable that regional language variation (but perhaps also regional variation in the folkloric way of expression) within the Seto region has contributed to the similarity and differences of word use, i.e. the geographic variation of language and culture may also have an effect on word use on smaller scales.

The ways of telling the stories and writing them down have been in constant evolution, as well as has been the language. Although the time period in question is short, from 1926 to 1953, it contains the remarkable turn in society caused by the events of the Second World War and Soviet occupation of Estonia from 1945. As folklore collections almost always provide the year of collection, we plotted this to find out if the date of collection has had any effect on the clustering of tales (Figure 13). As mentioned in the case of collectors, stories recorded by post-war collectors are placed near each other (white circles). We can also see that for most part the oldest recordings form distinct clusters at the edges of the graph. The central part with less idiomatic word use has gathered the stories from a slightly later period, which probably illustrates the fading of linguistic differences as well as storytelling tradition.

Along with fading of the storytelling tradition the stories on average become shorter. And it is important to reiterate here that stylometry expects the texts observed to be long enough for the expression of idiomatic style. Thus we may also suppose that the length of the texts has something to do with the similarity measures. From figure 14, where the graph of stories is identified by length of story in words, we can see that in general the central part of the graph – where the stories have a less distinct style – the stories tend to be in general shorter.

The aim of the current analysis was to find out if and to what extent the content and/or individual style of storytellers contributes to the similarity patterns of word form use.

The general idea of stylometry, that each author has his or her individual style of word use, seems to be confirmed. Stories told by the same performer clustered together more often while the classification of texts only rarely had an effect on the clustering. The analysis also revealed, however, notable effects of additional factors on the frequency distribution of the most common forms, i.e. length of text, individual style of collector, period of collecting, as well as the origins of the storytellers along with their local subdialects.

CONCLUSIONS

The analysis reveals the layered nature of geographical variation in a large corpus of Estonian runosongs showing how metric, stylometric, and typological variations follow different geographical patterns. The results of stylometric analysis, which detects the proximity of texts on the basis of the distribution of the most frequent words in regional runosong text collections are coherent in terms of geography, but relate also to linguistic, genre and content features of the texts. In order to explore further the application of stylometry to folklore texts, and how different aspects affect the results, we turned to the small regional corpus of fairy tales. The results of the stylometric analysis of fairy tales collected within the small Seto language community as a rule found the texts told by the same storytellers to be closer to each other than stories with the same content (tale types), although additional factors, especially the individual style of the collector, as well as the dialect variation within the Seto region, had a clear effect on the similarity of word use. The study also confirmed the principle of stylometry, which says that longer texts are in general able to reveal the style more clearly.

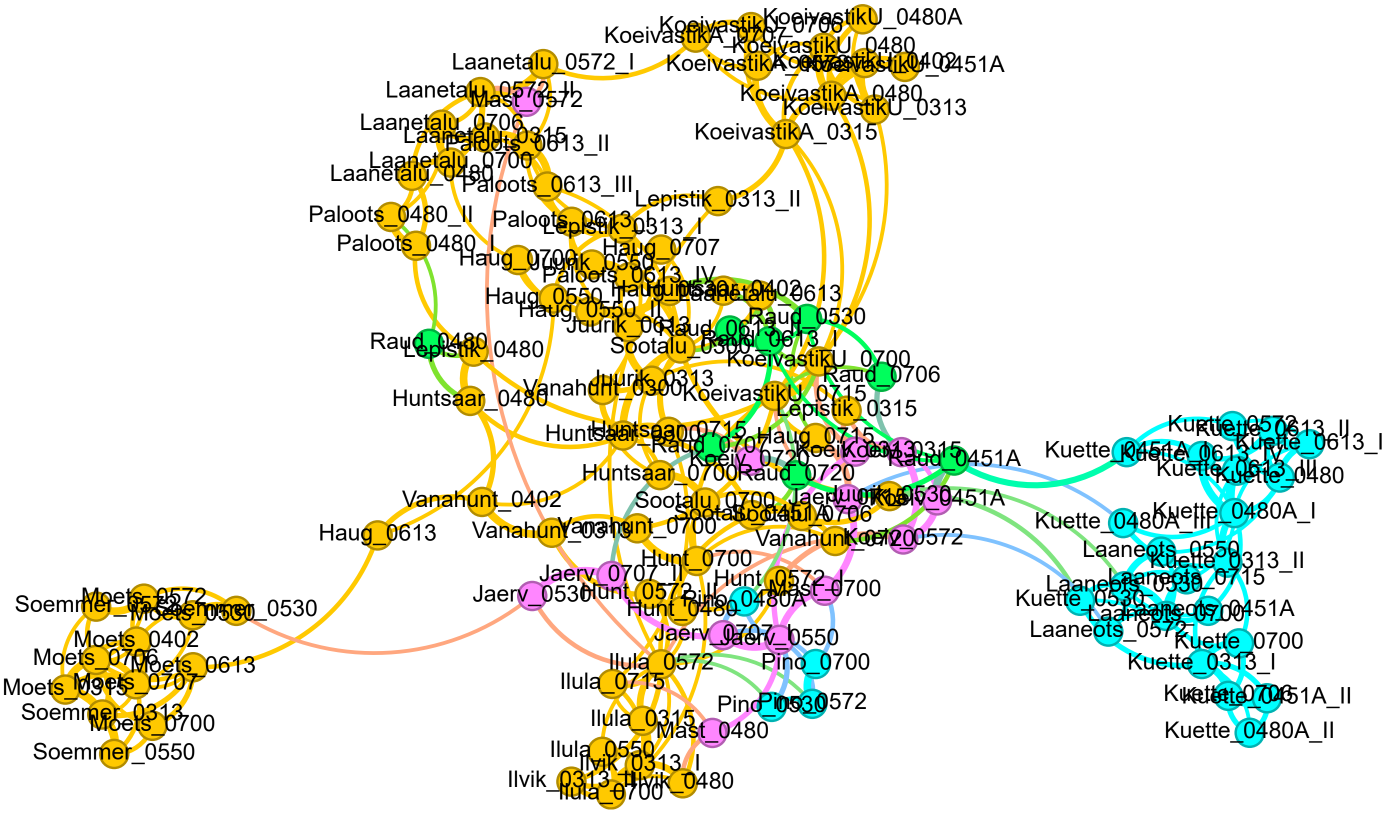

Figure 12. *Network of fairy tales, origin of storyteller identified by colour: Mäe community blue, Meremäe community pink, Saatse community green, Vilo community yellow.*

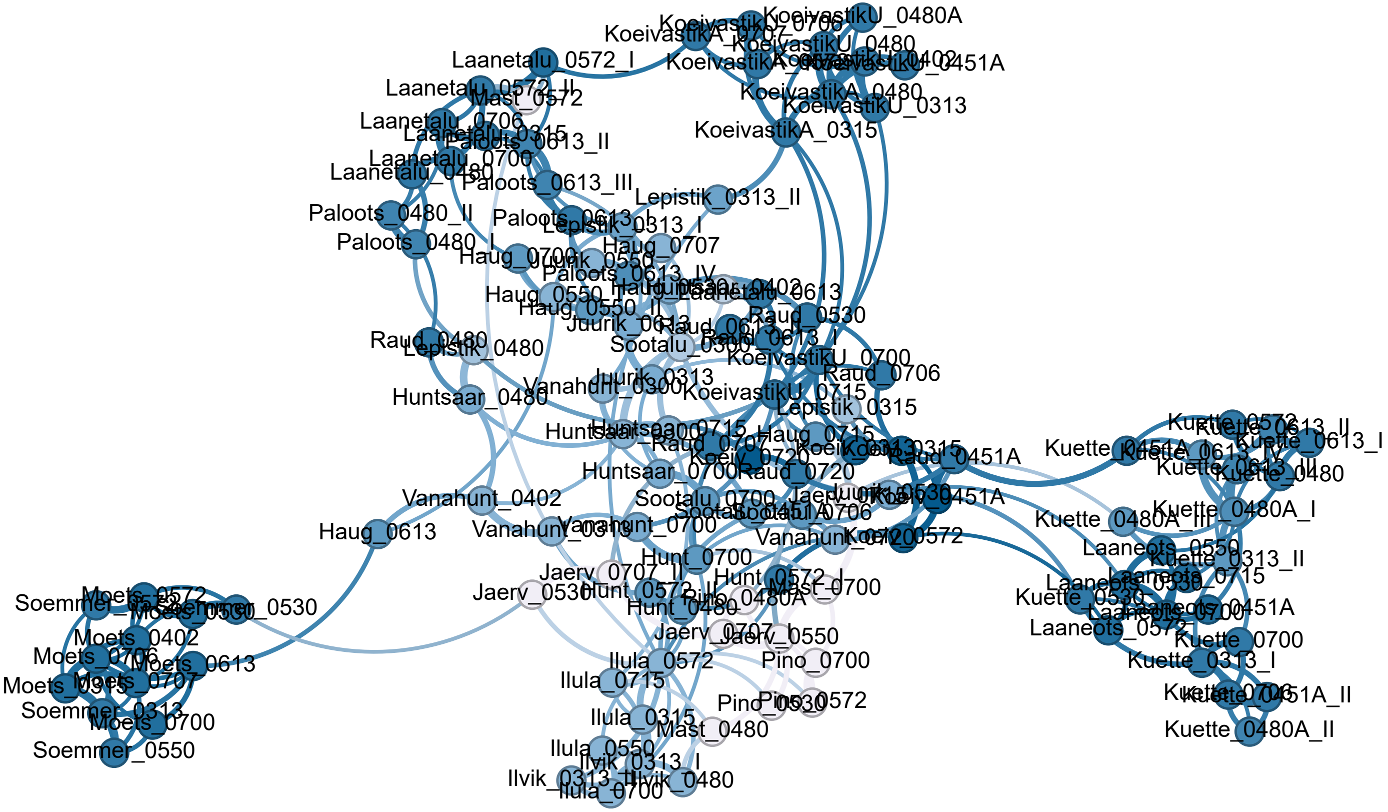

Figure 13. *Network of fairy tales, year of collection identified by colour from dark to light, i.e. from 1926 to 1953.*

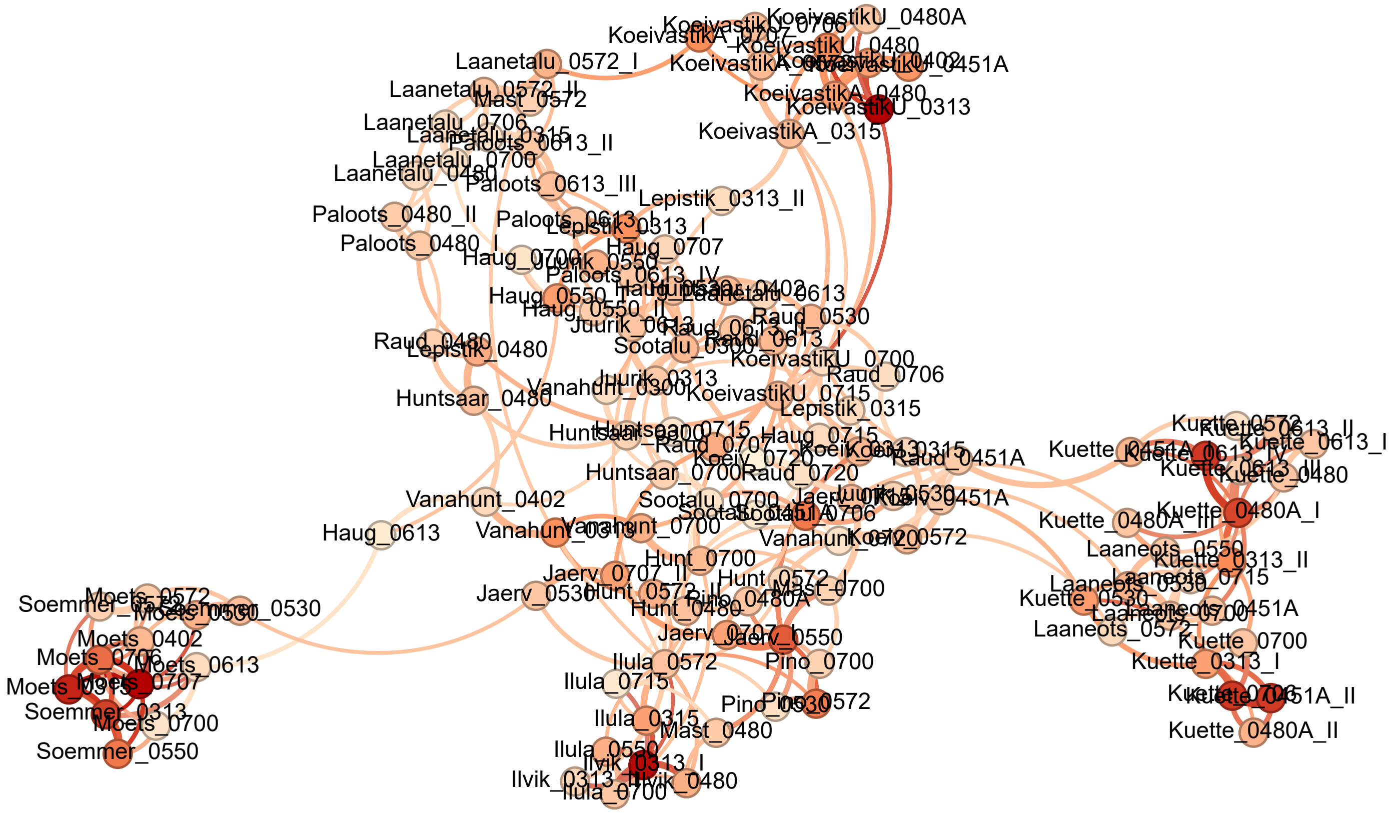

Figure 14. *Network of fairy tale text length identified by colour from light to dark, i.e. from 285 to 4,233 words.*

All in all, the study showed that textual variation is based on different factors that are not easily distinguishable. In the case of textual folklore, the folkloric variation is intertwined with linguistic variation. Regional and personal language usage influences folkloric expression through word choice, fixed formulae, alliterative constructions, etc, that design our way of thinking. An alliterative sentence, a line from the well-known runosong, *millal maksan memme vaeva* ‘when will I repay my mother’s love’, is considered one of the most beautiful sentences in Estonian. Rather than a final answer to any of our questions, we ask the reader instead if the beauty of this sentence is in its style or in its content?

ACKNOWLEDGEMENTS

This study was supported by the Estonian Research Council (PRG1288 A Corpus-based Approach to Folkloric Variation: Regional Styles, Thematic Networks, and Communicative Modes in Runosong Tradition), by the Estonian Ministry of Education and Research (EKKD65 Source Documents in the Cultural Process: Estonian Materials in the Collections and Databases of the Estonian Literary Museum II), by the Finnish Academy (project no. 333138 Formulaic Intertextuality, Thematic Networks and Poetic Variation across Regional Cultures of Finnic Oral Poetry), by the Estonian Literary Museum (project EKM 8-2/22/3 Folklore and its individual, community based, and institutional mechanisms of transmission) and by the European Regional Development Fund (TK 145 Centre of Excellence in Estonian Studies).

NOTES

[1] On the popular and academic terms for this tradition see Kallio et al. 2017.

[2] The issues of popularity and creativity, spread and collection density in archival folklore collections is thoroughly analysed in Krikmann 1997.

[3] We used the Gephi layout ForceAtlas2 by M. Jacomy, and filtered out the connections with negative and small values up to the value of 2.7 where every parish remained connected.

[4] On the impact of the cities on the development of Estonian dialects see Pajusalu 2013.

[5] For keyness analysis we used the R package quanteda (Benoit et al. 2018) and log-likelihood ratio method.

[6] Burrows (2002) estimates that the method works with texts of at least 1,500 words, approximately speaking.

[7] The Dragon-Slayer (ATU 300), The Magic Flight (ATU 313), The Faithless Sister (ATU 315), The Animal Bride (ATU 402), The Sister of Nine Brothers (AT 451A/Ee 451A), An Orphan and the Mistress's Daughter (ATU 480), Devil Wooing in the Sauna (ATU 480A), The Princess on the Glass Mountain (ATU 530), The Golden Bird (ATU 550), Skulls Making Noises (ATU 572*), The Rich Brother and the Poor Brother (ATU 613), Tom Thumb (ATU 700), The Maiden without Hands (ATU 706), The Miraculous Children (ATU 707), The Magic Cock (ATU 715), The Orphan as a Cuckoo (ATU 720), cf. Estonian tale summaries EMj I-1 2009: 589–615; EMj I-2 2014: 709–738.

[8] The stories of a third storyteller from Mäe community, Irina Pino, cluster together with the stories of other regions.

REFERENCES

Aarne, Antti 1910. *Verzeichnis der Märchentypen.* FF Communications 3. Helsinki: Academia Scientiarium Fennica.

Aarne, Antti & Thompson, Stith 1961. *The Types of the Folktale. A Classification and Bibliography. Second revision.* FF Communications 184. Helsinki: Academia Scientiarium Fennica.

Abello, James & Broadwell, Peter & Tangherlini, Timothy R. 2012. Computational Folkloristics. *Communications of the ACM* 55 (7), pp. 60–70. Available at http://dl.acm.org/citation.cfm?id=2209267. https://doi.org/10.1145/2209249.2209267.

Anderson, Walter 1923. *Kaiser und Abt. Die Geschichte eines Schwanks.* FF Communications 168. Helsinki: Academia Scientiarium Fennica.

Anderson, Walter 1935. *Studien zur Wortsilbenstatistik der älteren estnischen Volkslieder.* Eesti Rahvaluule Arhiivi toimetused 2. Tartu.

Anderson, Walter 1951. *Ein volkskundliches Experiment.* FF Communications 141. Helsinki: Suomalainen Tiedeakatemia.

Anderson, Walter 1956. *Eine neue Arbeit zur experimentellen Volkskunde.* FF Communications 168. Helsinki: Suomalainen Tiedeakatemia.

ATU = Uther 2004

Bastian, Mathieu & Heymann, Sébastien & Jacomy, Mathieu 2009. Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*, Vol. 8, pp. 361–362. San Jose, California, USA.

Bauman, Richard 1984. *Verbal Art as Performance*. Prospect Heights: Waveland Press.

Benoit, Kenneth & Watanabe, Kohei & Wang, Haiyan & Nulty, Paul & Obeng, Adam & Müller, Stefan & Matsuo, Akitaka 2018. quanteda: An R package for the Quantitative Analysis of Textual Data. *Journal of Open Source Software*, Vol. 3 (30), p. 774. https://doi.org/10.21105/joss.00774.

Beyer, Jürgen & Chesnutt, Michael 1997. Extracts from a Conversation with Isidor Levin. *Copenhagen Folklore Notes* 1–2, pp. 2–4.

Blondel, Vincent D. & Guillaume, Jean-Loup & Lambiotte, Renaud & Lefebvre, Etienne 2008. Fast unfolding of communities in large networks. – *Journal of Statistical Mechanics: Theory and Experiment*, No 10, P10008. https://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008.

Bogatyrev, Peter & Jakobson, Roman 1972 [1929]. Die Folklore als eine besondere Form des Schaffens. Blumensath, Heinz (Hg.). *Strukturalismus in der Literaturwissenschaft*. Köln: Kiepenheuer & Witsch, pp. 13–24.

Burrows, John 2002. ‘Delta’: A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, Vol. 17, No 3, pp. 267–287, https://doi.org/10.1093/llc/17.3.267.

de Certeau, Michel 1984 [1980]. *The Practice of Everyday Life*. Berkeley, Los Angeles, London: University of California Press.

Dorson, Richard M. 1963. Current Folklore Theories. *Current Anthropology*, Vol. 4 (1), pp. 93–112.

Dundes, Alan 1989. Interpreting “Little Red Riding Hood” Psychoanalytically. Alan Dundes (ed.) *Little Red Riding Hood: A Casebook*. Madison: The University of Wisconsin Press, pp. 192–236.

Eder, Maciej & Rybicki, Jan & Kestemont, Mike 2016. Stylometry with R: A Package for Computational Text Analysis. *R Journal*, Vol. 8 (1), pp. 107–121. https://journal.r-project.org/archive/2016/RJ-2016-007/.

Eder, Maciej & Górski, Rafal L. 2016. Historical Linguistics’ New Toys, or Stylometry Applied to the Study of Language Change. *DH 2016*, pp. 182–184.

Emj I-1 2009 = Järv, Risto & Kaasik, Mairi & Toomeos-Orglaan, Kärri (comps. and eds.) 2009. *Eesti muinasjutud* I:1. *Imemuinasjutud*. [Estonian Folktales I:1. Fairy Tales.] Monumenta Estoniae antiquae V. Tartu: Eesti Kirjandusmuuseumi Teaduskirjastus.

EMj I-2 2014 = Järv, Risto & Kaasik, Mairi & Toomeos-Orglaan, Kärri (comps. and eds.); Annom, Inge (ed.) 2014. *Eesti muinasjutud* I:2. *Imemuinasjutud*. [Estonian Folktales I:2. Fairy Tales.] Monumenta Estoniae antiquae V. Tartu: Eesti Kirjandusmuuseumi Teaduskirjastus.

ERAB 2023. *Eesti regilaulude andmebaas*. [The Estonian Runosong Database.] Compiled by J. Oras & L. Saarlo & M. Sarv. Tartu: Eesti Kirjandusmuuseumi Eesti Rahvaluule Arhiiv. Available at http://www.folklore.ee/regilaul/, last accessed on 25 March 2023.

Foley, John Miles 1985. *Oral-Formulaic Theory and Research: An Introduction and Annotated Bibliography*. New York.

Foley, John Miles 1992. Word-Power, Performance, and Tradition. *Journal of American Folklore*, Vol. 105, pp. 275–301.

Foley, John Miles 1995. *Singer of Tales in Performance.* Bloomington: Indiana University Press.

Hafstein, Valdimar R. 2001. Biological Metaphors in Folklore Theory. An Essay in the History of Ideas. *ARV. Nordic Yearbook of Folklore*, Vol. 57. Ed by Ulrika Wolf-Knuts. Uppsala: The Royal Gustavus Adolphus Academy, pp. 7–32.

Hagu, Paul & Pajusalu, Karl 2021. *Seto keele teejuht*. [A Guide to the Seto Language.] Värska: SA Seto Instituut.

Harvilahti, Lauri 1992. *Kertovan runon keinot. Inkeriläisen runoepiikan tuottamisesta*. [Devices of Narrative Poetry: Producing Ingrian Epic Poetry.] Helsinki: Suomalaisen Kirjallisuuden Seura.

Harvilahti, Lauri 2004. Vakiojaksot ja muuntelu kalevalaisessa epiikassa. [Stable Parts and Variation in Kalevala-metric Epic Folksongs.] In: Anna-Leena Siikala & Lauri Harvilahti & Senni Timonen (eds.). *Kalevala ja laulettu runo*. Helsinki: SKS, pp. 194–214.

Hiiemäe, Mall & Krikmann, Arvo 1992. On Stability and Variation on Type and Genre Level. *Folklore Processed. In Honour of Lauri Honko on his 60th Birthday 6th March 1992.* Studia Fennica Folkloristica 1. Edited by Reimund Kvideland. Helsinki: Suomalaisen Kirjallisuuden Seura, pp. 27–140.

Honko, Lauri 2000. Thick Corpus and Organic Variation: An Introduction. Lauri Honko (Comp.) *Thick Corpus, Organic Variation and Textuality in Oral Tradition.* Studia Fennica, Folkloristica 7. Helsinki: Finnish Literature Society, pp. 3–29.

Hymes, Dell 1981. *"In Vain I Tried to Tell You": Essays in Native American Ethnopoetics.* Philadelphia: University of Pennsylvania Press.

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*, pp. 96–100. Available at https://aclanthology.org/2022.nlp4dh-1.13/, last accessed on 3 October 2023.

Janicki, Maciej & Kallio, Kati & Sarv, Mari 2022. Exploring Finnic Written Oral Folk Poetry through String Similarity. *Digital Scholarship in the Humanities* [1–15]. https://doi.org/10.1093/llc/fqac034.

Järv, Risto & Sarv, Mari 2014. From Regular Archives to Digital Archives. In: Schmitt, Christoph (ed.). *Corpora Ethnographica Online. Strategies to Digitize Ethnographical Collections and Their Presentation on the Internet.* (Rostocker Studien zur Volkskunde und Kulturgeschichte; 5). Waxmann Verlag GmbH. pp. 49–60.

Järv, Risto 2016. The Singing Wolf to Meet His Kin Abroad. Web-based databases of the Estonian Folklore Archives. *Estudis de Literatura Oral Popular / Studies in Oral Folk Literature*, Vol. 5, pp. 29–44. https://doi.org/10.17345/elop201629-44.

Kaivola-Bregenhøj, Annika 2000. Varying folklore. *Thick corpus, organic variation and textuality in oral tradition.* Edited by Lauri Honko. Studia Fennica. Folkloristica 7 (NNF publications 7). Helsinki: Finnish Literature Society, pp. 93–130.

Kalkun, Andreas 2011. *Seto laul eesti folkloristika ajaloos. Lisandusi representatsiooniloole*. [Seto Singing Culture in Studies of Estonian Folklore: A Supplement to the History of Representation.] Dissertationes folkloristicae Universitatis Tartuensis 18. Tartu: Tartu Ülikool. Available at http://hdl.handle.net/10062/18222, last accessed on 3 October 2023.

Kalkun, Andreas 2015. Juhatusõst. [An Introduction.] *Ilosa' ja pogana' jutu'.* Seto Kirävara 10. Compiled and edited by Andreas Kalkun. Värska–Tartu: Seto Instituut, Eesti Kirjandusmuuseum, pp. 9–12.

Kallio, Kati & Frog & Sarv, Mari 2017. "What to Call the Poetic Form: Kalevala-Meter, Regivärss, Runosong, Alliterative Finnic Tetrameter, or Something Else?" *RMN Newsletter*, Vols. 12–13, pp. 139–161.

Kolk, Udo 1962. Värsisisesed vormelid eesti regivärsilises rahvalaulus. [Verse-internal Formulae in Estonian Runic Folksong.] *Tartu Ülikooli toimetised*, Vol. 117, pp. 71–155.

Krikmann, Arvo 1980. *Towards the typology of Estonian folklore regions. Paper presented to the Fifth International Finno-Ugric Congress, Turku, 1980.* Tallinn: Academy of Sciences of the Estonian S.S.R..

Krikmann, Arvo 1997. *Sissevaateid folkloori lühivormidesse*, I. [Insights Into Short Forms of Folklore I.] Tartu: Tartu Ülikooli kirjastus. Web-version: http://haldjas.folklore.ee/~kriku/LEX/KATUS.HTM.

Krikmann, Arvo 2014. *Uusi unistusi eesti murde- ja folkloorialade piiritlemise teemal.* [Some New Dreams on the Delineation of Estonian Dialect and Folklore Regions.] Tartu: EKM FO. Available at https://www.folklore.ee/~kriku/TRANSPORT/Geotypo.pdf, last accessed on 25 March 2023.

Krohn, Kaarle 1926. *Die folkloristische Arbeitsmethode. Begründet von Julius Krohn und weitergeführt von nordischen Forschern, erläutert von Kaarle Krohn*. Oslo [etc.]: H. Aschehoug.

Kuusi, Matti 1949. *Sampo-eepos. Typologinen analyysi.* [The Sampo Epic: A Typological Analysis.] Suomalais-Ugrilaisen Seuran Toimituksia 96. Helsinki.

Leino, Pentti 1970. *Strukturaalinen alkusointu suomessa.* [Structural Alliteration in Finnish.] Suomalaisen Kirjallisuuden Seuran Toimituksia 298. Helsinki: SKS.

Lindström, Liina & Pajusalu, Karl 2003. Corpus of Estonian Dialects and the Estonian Vowel System. *Linguistica Uralica*, Vol. 4, pp. 241–257.

Lord, Albert B. 1960. *The Singer of Tales.* Cambridge, MA: Harvard University Press.

Lotman, Jurij [Juri] 1977 [1970]. *The Structure of the Artistic Text.* Michigan Slavic Contributions 7. University of Michigan.

Mäkinen, Martti 2020. Stylo Visualisations of Middle English Documents. *Journal of Data Mining & Digital Humanities*, pp. 1–10. Available at https://jdmdh.episciences.org/7022/pdf, last accessed on 25 March 2023.

Normann, Erna 1935. Kurg kündmas. [The Stork Plowing.] *Kaleviste mailt*. Õpetatud Eesti Seltsi Kirjad 3. Tartu: Õpetatud Eesti Selts, pp. 34–37.

Ong, Walter J. 1982. *Orality and Literacy: The Technologizing of the Word*. London and New York: Methuen.

Pajusalu, Karl 2013. Eesti keeleala piirid. [The Borders of the Estonian Language Area.] *Keel ja Kirjandus*, Vol. 3, pp. 210–213.

Parry, Milman 1930. *Studies in the Epic Technique of Oral Verse-Making I: Homer and Homeric Style*. Harvard Studies in Classical Philology 41, pp. 73–148.

Propp, V[ladimir] 1968 [1928]. *Morphology of the Folktale* (2nd ed., American Folklore Society Bibliographical and Special Series, Vol. 9). Austin: University of Texas Press.

Pöysä, Jyrki 2000. Variation in Archived Anecdotes. *Thick Corpus, Organic Variation and Textuality in Oral Tradition.* Studia Fennica Folkloristica 7. Edited by Lauri Honko. Helsinki: Finnish Literature Society, pp. 577–593.

Reichl, Karl 2007. *Edige. A Karakalpak Oral Epic as Performed by Jumabay Bazarov.* FF Communications 293. Helsinki: Suomalainen Tiedeakatemia.

Rüütel, Ingrid 2012. *Eesti uuema rahvalaulu kujunemine*. [The Formation of the Modern Estonian Folk Song.] Tartu: Eesti Kirjandusmuuseum.

Sadeniemi, Matti 1951. *Die Metrik des Kalevala-Verses.* FF Communications 139. Helsinki.

Salve, Kristi & Sarv, Vaike 1987. *Setu lauludega muinasjutud.* [Seto Fairy Tales with Songs.] Tallinn: Eesti Raamat.

Sarv, Mari 2000. *Regilaul kui poeetiline süsteem.* [Runosong as a Poetic System.] Paar sammukest XVII. Eesti Kirjandusmuuseumi aastaraamat. Tartu: Eesti Kirjandusmuuseum.

Sarv, Mari 2008. *Loomiseks loodud: Regivärsimõõt traditsiooniprotsessis.* [Created for Creation: The Verse Metre of Estonian Regilaul in the Tradition Process.] Eesti Rahvaluule Arhiivi toimetused 26. Tartu: Eesti Kirjandusmuuseumi Teaduskirjastus.

Sarv, Mari 2015. Regional Variation in Folkloric Meter: The Case of Estonian Runosong. *RMN Newsletter*, Vol. 9, pp. 6–17.

Sarv, Mari 2019. Poetic Metre as a Function of Language: Linguistic Grounds for Metrical Variation in Estonian Runosongs. *Studia Metrica et Poetica*, Vol. 6 (2), pp. 102–148. https://doi.org/10.12697/smp.2019.6.2.04.

Sarv, Mari & Oras, Janika 2020. From Tradition to Data: The Case of Estonian Runosong. *ARV. Nordic Yearbook of Folklore*, Vol. 76, pp. 105–117.

Seljamaa, Elo-Hanna 2005. Walter Anderson: A Scientist beyond Historic and Geographic Borders. In: K. Kuutma & T. Jaago (eds.). *Studies in Estonian Folkloristics and Ethnology. A Reader and Reflexive History.* Tartu: Tartu University Press, pp. 153–168.

SKVR 2021. *SKVR-tietokanta – kalevalaisten runojen tietokanta*. [The SKVR Database: A Database of Kalevalaic Poems.] Helsinki: Suomalaisen Kirjallisuuden Seura. Available at http://skvr.fi, last accessed on March 25, 2023.

SNE 2014 = *Seto naisi elolaulu'. Antoloogia*. [Life Songs by Seto Women.] Seto Kirävara 8. Compiled by Andreas Kalkun & Vahur Aabrams. Värska–Tartu: Seto Instituut, Eesti Kirjandusmuuseum, pp. 271–309.

Steinitz, Wolfgang 1934. *Der Parallelismus in der Finnisch-Karelischen Volksdichtung*. FF Communications 115. Helsinki.

Sykäri, Venla 2014. Improvisation as a Singer's Concept in Oral Poetry. In: Huttu-Hiltunen, Pekka & Frog & Lukin, Karina & Stepanova, Eila (eds.). *Song and Emergent Poetics: Laulu ja runo*. Kuhmo: Juminkeko, pp 91–98.

Šeļa, Artjoms 2021. Erinevused, kaugused ja sõrmejäljed. Stilomeetria ja mitmemõõtmelise tekstianalüüsi alused. [Differences, Distances and Fingerprints: Stylometry and the Foundations of Multidimensional Text Analysis.] *Keel ja Kirjandus*, Vol. 64 (8–9), pp. 696–718. https://doi.org/10.54013/kk764a3.

Tampere, Herbert 1932. Laul jänese õhkamisest eesti rahvatraditsioonis ja kirjanduses. [A Song about a Hare Sighing in Estonian Folk Tradition and Literature.] In: *Vanavara vallast.* Õpetatud Eesti Seltsi Kirjad 1. Tartu: Õpetatud Eesti Selts, pp. 59–116.

Tangherlini TR, Shahsavari S, Shahbazi B, Ebrahimzadeh E, Roychowdhury V (2020) An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, Pizzagate and storytelling on the web. *PLoS ONE* 15(6): e0233879. https://doi.org/10.1371/journal.pone.0233879.

Tedlock, Dennis 1983. *Spoken Word and the Work of Interpretation*. Philadelphia: University of Pennsylvania Press.

Tedre, Ülo 1964. Stereotüüpsusest Karksi rahvalauludes. [On Stereotypy in Karksi Folksongs.] *Eesti rahvaluulest.* Tallinn, pp. 52–86.

Uther, Hans-Jörg 2004. *The Types of International Folktales: A Classification and Bibliography. Based on the system of Antti Aarne and Stith Thompson.* FF Communications 284–286. Helsinki: Suomalainen Tiedeakatemia.

Wolf-Knuts, Ulrika 2000. On the History of Comparision in Folklore Studies. In: Lauri Honko (ed.). *Thick Corpus, Organic Variation and Textuality in Oral Tradition*. Studia Fennica. Folkloristica 7. Helsinki: Finnish Literature Society, pp. 255–283.

Mari Sarv is research professor at the Estonian Folklore Archives of the Estonian Literary Museum. Her main field of research is Estonian and Finnic runosong, especially its language and poetics. She has published and edited books on the topic, organized conferences, and contributed to the creation and development of the database of Estonian runosongs. Benefitting in her research from computational methods herself, she has been also propagating the ideas of data availability and use of computational methods in humanities research,

initiating Estonian digital humanities conference series in 2013. She has been leading several research projects on topics related to folklore archives, runosongs, and cultural heritage.

mari@haldjas.folklore.ee

Risto Järv is the head of the Estonian Folklore Archives of the Estonian Literary Museum, and associate professor on Estonian and comparative folklore at the University of Tartu. His main subject of interest is the genre of fairy tales. Together with his work group he is managing the database of Estonian fairy tales as well as publishing the series of academic editions on the subject. His research interests cover also historiography of Estonian folkloristics, proverbs, and contemporary re-interpretations of narrative folklore.

risto.jarv@folklore.ee