Exploring Finnic written oral folk poetry through string similarity

Maciej Janicki
Department of Digital Humanities, University of Helsinki, Finland

Kati Kallio
Finnish Literature Society, Finland

Mari Sarv
Estonian Literary Museum, Estonia

Abstract

Suomen Kansan Vanhat Runot (*Old Poems of the Finnish People*) is a collection of nearly 90,000 oral folk poems written down between 1564 and the early 20th century. It is characterized by frequent reoccurrence of similar pieces of text on various levels (from entire poems, through passages to single verses and collocations). However, finding these similarities is challenging due to a high degree of orthographical, morphological, and compositional variation. In this article, we propose a method for automatically identifying *equivalent verses*, i.e. verses conveying the same meaning with the same words, using a clustering based on cosine similarity of character bigram vectors. The method achieves around 81% F-score and has been successfully used for identifying similarities across the entire SKVR corpus on the level of verse, passage, and poem. The results can be browsed through a Web interface.

Correspondence:
Maciej Janicki, Department of Digital Humanities, University of Helsinki, Unioninkatu 40, 00170 Helsinki, Finland.
E-mail: maciej.janicki@helsinki.fi

1 Introduction

The collection *Suomen Kansan Vanhat Runot* (*Old Poems of the Finnish People*)[1] consists of nearly 90,000 oral folk poems written down by various collectors between 1564 and the early 20th century. It contains archaic epics, lyrical songs, ballads, ritual songs, charms, dancing songs, lullabies, and many other genres in similar poetic forms and four Finnic languages: Karelian, Izhorian, Votic, and Finnish. The corpus is an important source in analysing e.g. historical oral and popular culture, verbal art, belief systems, and mythology (see Kallio *et al.*, 2017). Most of the poems were recorded by Finnish scholars during the 19th and early 20th century, when Finnish was a minority language of the autonomous Russian Grand Duchy of Finland (1809–1917). Finnic folk poetry played a great role in developing Finnish and Karelian literary languages and identities. The national epic Kalevala is a literary work Elias Lönnrot created using oral poems he and others had collected in Karelia, Ingria, and Finland (Haapoja-Mäkelä *et al.*, 2018).

The SKVR corpus has certain unique properties that make the use of computational methods both challenging and rewarding. Just the size of the corpus—nearly 90,000 poems totaling over 1.4 million verses—already suggests the use of digital searches. Typically, oral poetry is extremely diverse: the degree and character of variation itself varies according to local singing culture, singer, genre of poem, and the performance situation. According to the oral-formulaic theory (see Lord, 1960; Foley, 1988) oral poems were not learnt by heart, but the singers

https://doi.org/10.1093/llc/fqac034

mastered the traditional poetic language and metre, common formulas (recurring poetically motivated collocations), motifs, story-lines, performance styles, and ways to combine and use these to create and re-create versions for different purposes and social settings. Finnic tradition contains both highly variable and stable examples (see e.g. Harvilahti, 1992a; Timonen, 2004). Thus, some poems were recorded in several nearly-identical local versions, while sometimes one story (or poem type) may be told with very different motifs and lines, or new poems created on the basis of traditional poetic language, formulas, and motifs. This leads to some verses and motifs being typical to particular poem types, some verses and motifs appearing in very different poetic contexts and meanings, and poem types often appearing in all kinds of combinations or even cycles. Thus, similar formulas, verses, and short sequences of verses recur in different poetic context across the dataset, often in different dialectal, morphological, and orthographic forms. In a typical research setting analysing the corpus or some substantial part of it, the central question is how to find all the relevant instances of one poem, motif, or verse type. This was taken into account when the original book series (1908–48) was published: each volume includes a type index, which were later unified and re-analysed into a single type index. However, such indexing of extremely diverse material is inevitably subjective, and the categories are often ambiguous. In addition, most of the index works at the level of poem types, not of smaller motifs let alone individual verses, which scholars are however often interested in analysing. At the level of the entire corpus, we do not know how the verse-level intertextuality between different poem types and motifs works, or how different singers, poem types or regional singing traditions relate to one another in terms of making use of similar motifs, verse types, and formulas. Automatic detection of similarities would significantly facilitate the selection of relevant subcorpora for qualitative research, recognizing significant features in the corpus, and also potentially enabling some computational analysis.

However, the similarities of the content units (lines, motifs, types) in the corpus are obfuscated by a remarkable amount of linguistic and orthographic variation, which makes the computational detection of similarities a non-trivial task. Table 1 illustrates the variation of the verse type *Savu saarella palaa*, ('Smoke is burning on the isle'). The list is not exhaustive—a total of seventy different forms could be automatically identified using the method introduced in this article.

Table 1. Some variants of the verse type *Savu saarella palaa* ('Smoke is burning on the isle,'), with place of occurrence (SKVR volume and poem number)

Savu saarella palavi,	I1 163 a)
Savu suaressa palaubi,	I2 702
Mi se savu soarella palavi,	I2 705
Savu soaressa palauve,	I2 702
Savu saarella palaa,	I4 753
Savu suaressa palaa,	II 206 a)
Šavu šoaressa palaubi,	II 208
Savu soarella palaabi,	VII1 781
Savu soarell' on palaabi,	VII1 808
Savu soarella palaa,	VII1 821 a.
Savu saarell' on paloa	VII1 781 d.
Savubon soarella palaabi,	VII1 784
Savut saarella palavat:	VII1 790
Savupa soarelle palaabi,	VII1 803

The examples differ with respect to phonology or orthography (*s* ∼ *š*; *aa* ∼ *oa* ∼ *ua*), morphology (case ending *-lla* ∼ *-lle* ∼*-ssa*; 3rd person singular verb ending: *-V* ^{2} ∼ *-Vbi* ∼ *-ubi* ∼ *-vi* ∼ *-uve* etc.), as well as insertion of filler words (*on* 'is', *se* 'it') and particles (*-pa*) and shortening of endings (*-ll'*), yet all of them have the same structure and convey the same meaning using the same wording. We subsequently call such verses *equivalent* and aim to find all equivalent instances of a verse in the corpus. As the verses in this Finnic poetic tradition typically consist of 7–10 syllables, which amounts to 2–3 words on average, equivalent verses oftentimes do not have a single word in common that would be spelled exactly the same. This makes full-text search challenging, especially when the variants are difficult to anticipate.

In this article, we propose an unsupervised computational method for identifying clusters of equivalent verses. Further, we show how the results can be used to automatically find parallel variants of the same poem type or passage. We present the method for finding equivalent verse pairs in Section 3 and evaluate it in Section 4. In Section 5 we present a prototype Web interface *Runoregi*, which allows users to interactively explore the automatically computed similarity on various text levels (verse, poem, passage).

2 Related Work

2.1 Variation and similarity in oral tradition

Ever since the beginning of folklore studies, the central focus has been on variation and similarity, change and continuities at different levels of oral traditions (see Bendix and Hasan-Rokem, 2012). In the field of oral poetry, a pivotal turning point was the work of Milman Parry and Albert Lord on Serbo-Croatian epics: they understood the creative way in which the singers built the poems in performance on the basis of traditional poetics, formulas, themes, and storylines (Lord, 1960). Since then, oral-formulaic theory has been tested and developed, making evident that the characteristics, levels and degree of variation vary a lot according to tradition, and also relate, as Lord noted, to the length of the songs (see Foley, 1988; Harvilahti, 1992a, 185n10). Formula was defined by (Parry 1930, p. 80) as 'a group of words which is regularly employed under the same metrical conditions to express a given idea'. The first one to apply this approach fully to the Finnic oral tradition was Lauri Harvilahti (1992a, 1992b, 2000; see also Kuusi, 1967), who used manual lemmatization to explore the formulaic system of Ingrian oral poems computationally, as well as testing the applicability of different views on oral formulaic theories to this tradition. Liina Saarlo (e.g. 2005) has studied formula systems in Estonian oral poetry, and Tiiu Jaago (2016), Lotte Tarkka (2013), and Frog (2014, 2016) have applied the concept to various Finnic traditions as well (see also Frog and Lamb, 2022). In addition, the research on poetic metre and language (see Sarv, 2008, 2015, 2019; Saarinen, 2018, for overviews), and of parallelism (see Frog and Tarkka, 2017; Sarv, 2017) often relates to similar questions. In these studies, the number of poems and verses analysed manually may be substantial—e.g. Matti Kuusi (1949; see 1990, p. 136) examined 41,762 lines by hand and Lauri Harvilahti (1992a) lemmatized 22,333 lines. Research concentrating on e.g. genre systems, intertextuality, or performance often also analyses the structures of songs (e.g. Arukask, 2003; Tarkka, 2013; Timonen, 2004). An earlier study by Mari Sarv (2004) has analysed the statistical distribution of singers' verse repertories geographically as well as in terms of stereotypy and uniqueness on the basis of songs from a distinct region in North-East Estonia.

Over recent decades, there has been growing interest in computational folkloristics (see Abello *et al.*, 2012; Harvilahti, 2019; Tangherlini, 2016) as the historical sources are being digitized (see Ilyefalvi, 2018), or in analysing contemporary culture (see Hakamies and Heimo, 2019). However, the complex folkloric and linguistic variation along with the fluctuations in collection history poses a considerable challenge for finding appropriate methods for large-scale analysis.

2.2 Text reuse in Digital Humanities

The computational task of recognizing similar text lines and passages in large collections of texts occurs in Digital Humanities in various scholarly contexts and is commonly referred to as *text reuse*. For example, in the study of classical corpora, scholars are often interested in finding literary influences, as well as citations and paraphrases of influential thinkers, which—contrary to the modern tradition—were not explicitly marked as such. Furthermore, texts existing in multiple editions are sometimes aligned in order to study their similarities and differences.

Methods for text reuse detection are often multi-staged and complex in technical detail (see Büchler (2013) for a comprehensive and systematic overview). In a simplified view, a common approach seems to be to initially detect candidates for similarity by searching for common words or word *n*-grams between documents. Subsequently, such matches can be scored based on a more detailed and more computationally intensive comparison of their surrounding context (e.g. alignment). Examples of this workflow include Olsen *et al.* (2011) (tracking sources of an 18th century encyclopaedia), Smith *et al.* (2014) (US legislation and 19th century newspapers), and Sturgeon (2018) (classical Chinese texts; using characters rather than words). Shmidman *et al.* (2018) use a similar method but address the problem of orthographic variation in the Babylonian Talmud by reducing each word to its two least frequent letters.

The same general workflow has been applied to poetry in the Tesserae project (Coffee *et al.*, 2013). Originally developed for studying Latin poetry, it has been applied to quantitative studies (Bernstein *et al.*, 2015) and an attempt to adapt the method to

English texts has recently been made by Shang and Underwood (2021).

When dealing with versified poetry, similarity between individual verse lines provides a good starting point for comparing longer passages. Jänicke and Wrisley (2017) align and visualize different editions of Medieval French poems using a combination of word *n*-gram overlap and relative edit distance for detecting similar lines. This method is able to address the problem of linguistic and orthographic variation, which is often encountered in historical texts. Meinecke *et al.* (2019) modified the method further to include similarity based on word embeddings.

Contrary to poetry, large volumes of continuous text provide additional difficulties, as there is no natural structural unit (like the line) to base the comparison on. In such cases, improvements have recently been achieved by repurposing the BLAST algorithm, originally designed for comparing biological sequences, for text reuse (Vesanto *et al.*, 2017; Vierthaler and Gelein, 2019). Furthermore, Broadwell *et al.* (2017) presented a method for identifying similar passages of text based on hashing sequences of consecutive words using a technique called Locality Sensitive Hashing, which efficiently recognizes similar sequences. The method was successfully applied to a collection of Danish folk legends recorded from oral tradition.

3 Identifying equivalent verses

We begin the similarity search by finding similar verses (i.e. lines) across the entire corpus, without even taking into account the division into documents (poems). This approach utilizes an important property of metric poetry: a similar (longer) passage of text will always be segmented into verses in the same way. In particular, it has been pointed out that verses in Finnic oral poetry are self-contained linguistic units and verse boundaries typically correspond to sentence or syntactic constituent boundaries (Leino, 1975). Thus, any longer-range similarity remains recognizable when comparing individual verses. Given this natural unit of text segmentation, the task becomes much easier than finding similar passages in continuous text would be.

Drawing on the example shown in Table 1, we aim to identify *equivalent verses*, i.e. verses that consist of the same content words, despite differing in their string form. This will be achieved in two steps: (1) finding pairs of similar verses and (2) clustering the graph of similarities to identify sets of equivalent verses.

3.1 Verse similarity

For the first step, due to the size of the corpus (over 1.4 million verses), an efficient way of finding the nearest neighbours to a verse without needing to compare it against all other verses is desirable. The most common approach for this kind of similarity search is to represent the object of interest (here: a verse) as a vector in a high-dimensional space. For such representation, efficient algorithms for nearest neighbour search are readily available.

A simple and effective numeric representation is the *bag-of-bigrams*: each verse is represented by a vector of character bigram frequencies (Table 2). Prior to that, the strings are lowercased and digits and punctuation marks are removed. The number of dimensions is restricted to d most frequent bigrams across the whole corpus, with values around $d = 300$ being considered sufficient (see Table 4 in Section 4).[3]

We consider bigrams to be the optimal tradeoff between granularity and the number of dimensions—trigrams would generate a prohibitive number of dimensions,[4] while also missing some similarities, whereas unigrams readily produce high similarities for unrelated verses.[5] Somewhat surprisingly, no verse pairs with highly similar bigram vectors and completely different content were observed. Thus, the bigrams preserve enough information about the arrangement of letters in the verse.

As a similarity metric for verse pairs, we use cosine similarity[6] of bigram vectors. For two vectors $\mathbf{x}, \mathbf{y}$, the cosine similarity (geometrically representing the

Table 2. Vectorization of preprocessed verses into bigram frequencies

	sa	so	av	vu	u_	_s	oa	ar	el	ll	la	es	ss	...
savu_saarella_palavi	2	0	2	1	1	1	0	1	1	1	2	0	0	
savu_soaressa_palauve	2	1	1	1	1	1	1	1	0	0	1	1	1	
savu_suaressa_palaubi	2	0	1	1	1	1	0	1	0	0	1	1	1	

cosine of the angle between the vectors) is given by the formula:

$$\cos\angle(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{||\mathbf{x}||||\mathbf{y}||} = \frac{\sum_{i=1}^{d} x_i y_i}{\sqrt{\sum_{i=1}^{d} x_i^2}\sqrt{\sum_{i=1}^{d} y_i^2}}. \quad (1)$$

Cosine similarity measures whether the vectors point in the same direction, which for our representation translates to the similarity in *proportions* of the particular bigrams, rather than their absolute frequencies. For example, hypothetical verses with duplicated content, like '*Savu savu saarella saarella palaavi palaavi*' or '*Savu saarella palaavi savu saarella palaavi*', would have the same similarity scores to other verses as '*Savu saarella palaavi*'. This is not a practical problem because verses where the entire content is repeated do not exist and, for the vast part of the corpus, the verse length is approximately constant, conforming to the eight verse positions of poetic metre. Even repetitions of a single word are very rare.

Also, the insertion of additional short words has little effect on the direction of the vector. This is desirable because equivalent verses might differ in terms of short function words (see: '*Mi se savu soarella palavi*' in Table 1). Finally, the information about the ordering of the bigrams (and thus words) in a verse is lost during vectorization, so verses with transposed word order (like the former and the hypothetical but nonexistent '*Palaavi saarella savu*'), are expected to have high similarity. This, again, is desirable.

3.2 Similarity computation

For each of the 1.4 million verses from the corpus, we retrieve up to $k = 1{,}000$ nearest neighbours with cosine similarity greater than some threshold α. The optimal threshold value turned out to be between 0.7 and 0.8 (see Section 4), as below 0.7 false positives start to appear in large masses. The similarity computation can be implemented efficiently using the FAISS library[7] (Johnson *et al.*, 2017) which implements advanced algebraic methods for indexing large numbers of vectors, allowing quick retrieval of nearest neighbours. FAISS includes a GPU implementation that is able to carry out this computation in around an hour on a mid-range desktop PC equipped with a GeForce GTX 1060 GPU. An example of nearest neighbour computation results is shown in Table 3.

Table 3. Selected nearest neighbours for *Savu saarella palaa*, (not all shown)

Savu saar[ella] p[alaa],	1
Savu saarella palaavi,	0.964396
Savu saarella palaapi,	0.960769
Savu saarella palavi,	0.938971
Savu saarella palapi.	0.936382
Savu soarella palaa,	0.912871
Savu soarella palaavi,	0.875
Savu soarella palaabi,	0.870388
Savut saarella palavat:	0.864242
Savu suarella palavi,	0.808694
Savu soarella palaupi,	0.78335

Table 4. Results of the similarity+clustering method

d	α	Precision	Recall	F-score
100	0.6	49.0±0.8%	97.0±0.7%	65.1±0.8%
	0.7	58.4±0.7%	94.2±0.6%	72.1±0.7%
	0.75	64.4±0.5%	91.0±0.6%	75.4±0.3%
	0.8	71.6±0.6%	86.3±1.5%	78.2±0.8%
	0.85	81.9±0.9%	74.4±2.8%	77.9±1.8%
	0.9	91.4±0.3%	57.1±1.7%	70.3±1.3%
200	0.6	54.1±0.6%	98.7±0.4%	69.9±0.5%
	0.7	63.7±0.5%	95.7±0.9%	76.5±0.4%
	0.75	69.4±0.3%	91.5±1.1%	78.9±0.5%
	0.8	76.8±0.6%	85.9±0.4%	**81.1**±0.5%
	0.85	84.8±0.8%	71.4±2.7%	77.5±1.7%
	0.9	92.9±0.3%	50.5±1.2%	65.4±1.0%
300	0.6	55.2±0.3%	98.7±0.2%	70.8±0.2%
	0.7	64.8±0.8%	94.3±0.5%	76.8±0.5%
	0.75	70.2±0.8%	90.1±1.5%	78.9±1.0%
	0.8	78.0±0.7%	84.7±0.9%	**81.2**±0.6%
	0.85	85.9±0.7%	70.4±2.5%	77.4±1.7%
	0.9	93.6±0.2%	49.9±1.4%	65.1±1.2%

In addition to speed and scalability, a further reason speaking in favor of the bag-of-bigrams method is its high recall: given a low-enough threshold, it successfully finds different kinds of similarity, including half-verse correspondences, alliteration patterns, similar-sounding verses (also without similarity in meaning) and more. Even if we aim for narrower criteria and higher precision, this method can always be used for finding candidate pairs that are passed for further filtering.

3.3 Clustering

Although the nearest-neighbour lists already provide good pairs of equivalent verses, they also capture similar but non-equivalent types. For example, for the verse *Eipä löyä Väinämöistä*, the nearest neighbour search finds both *Eibä* ***löüvä*** *Väinämöistä* (dialectal variation), as well as *Eipä* ***sopin*** *Väinämöistä* (different verb). These cases cannot be separated simply by a general threshold on similarity.

Moreover, the nearest neighbour relation is not transitive: if A is a neighbour of B and B is a neighbour of C, A is not necessarily a neighbour of C. In order to speak of *equivalent* verses, it is desirable to define an equivalence relation in the mathematical sense, thus a transitive one. This would allow us to map verses to a set of abstract 'verse types', in which equivalent verses would be assigned the same type. Among others, such mapping is necessary for the 'bag-of-verses' method for computing similarity between entire poems, which is presented in Section 5.

The mapping of verses to equivalence classes will be achieved by a graph clustering algorithm. The results of the similarity computation can be represented as a graph, in which verses are nodes and pairs of verses with a similarity exceeding a certain threshold are connected with an edge. Sets of equivalent verses, like the one in Table 1, are expected to form densely connected clusters in this graph, as each of the verses will be similar to most others.[8] On the other hand, similar but non-equivalent verses will be more weakly and sparsely connected, and thus will not form clusters. The graph of similarities is weighted, with weights being computed as follows: if ξ is the cosine similarity between two verses, with $\xi \in (\alpha, 1]$, the edge between those verses is weighted with $w = \frac{\xi - \alpha}{1 - \alpha}$, so that the range of possible values of w is scaled to the interval $(0, 1]$.

For clustering we apply the Chinese Whispers algorithm (Biemann, 2006). It is chosen mainly due to its simplicity and linear runtime wrt. the number of graph edges, which is essential given the size of our graph. This computation takes around 15 min on the aforementioned hardware. The clustering is illustrated in Fig. 1.

4 Evaluation

The key notion in the methods presented in the previous section is *equivalent* verses: a pair of verses that are considered 'the same' on some abstract level, while possibly differing in their string form. Because it is an equivalence relation in the mathematical sense, i.e. reflexive, symmetrical and transitive, the clustering

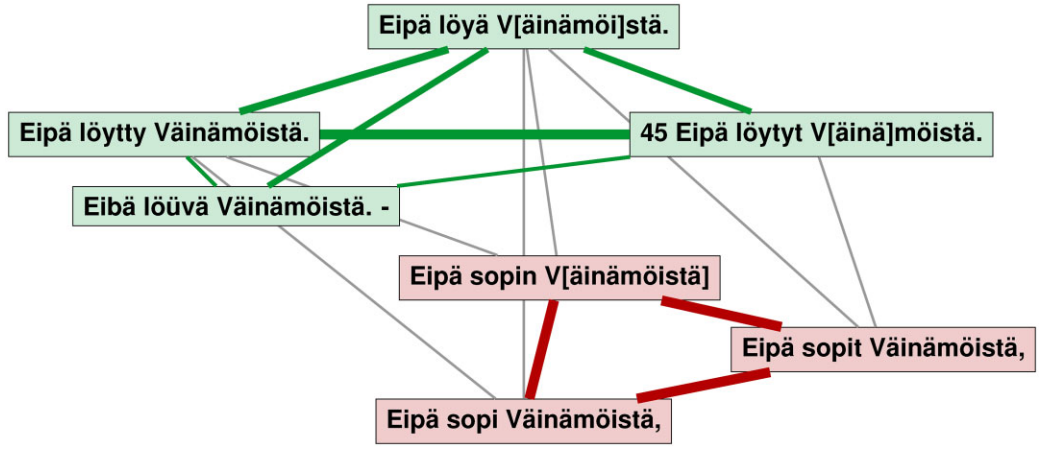


Fig. 1 The Chinese Whispers algorithm finds the most densely connected clusters within the graph of verse similarities. (Edge thickness corresponds to weights, colour to the resulting clusters)

divides the corpus into *equivalence classes*, inside which every verse is equivalent to every other. The examples listed in Table 1 are part of such a class.

4.1 Dataset annotation

To evaluate the methods, we formulated criteria for a verse pair to be considered equivalent from the point of view of oral poetry and folklore studies. Depending on the research interests, the equivalence of a verse pair or poetic formula may have very different definitions (see e.g. Foley, 1988; Harvilahti, 1992a; Sarv, 2004; Saarinen, 2018; Frog and Lamb, 2022). Here, we emphasized similar content words. An equivalent verse pair includes the same content words (nouns, adjectives, verbs). These may also be different words deriving from the same word stem (*päistäre* ∼ *päistärikkö*), whereas the function words (such as prepositions, pronouns, and conjunctions) and interjections may differ. Different word order and question-answer or affirmation-negation pairs are allowed, as is orthographic, morphological, and dialectal variation. Use of synonyms (such as *porsas* 'pig, piglet, pork' for *sika* 'pig, swine, boar, pork') is not considered equivalent. Cases that would necessitate checking the verse in poetic context (e.g. non-conventional abbreviations) are defined as non-equivalent. This definition allows some difference in the meaning and corresponds to one typical level of oral variation. In some other contexts, it would be highly relevant to search for verses using synonyms or words that sound rather than mean the same, or verses that share only some similar content words.

We subsequently prepared a data sample for manual annotation. Sampling verse pairs presented a challenge, because on one hand the sampling method should be as independent as possible from the approach that is being evaluated, but on the other hand the sample should contain a balanced number of equivalent and non-equivalent pairs. This would be impossible to achieve when choosing the pairs with uniform probability.

In order to guarantee a sufficient amount of equivalent pairs, we first restricted the evaluation corpus to six poem types that are common across the different languages and dialects of the corpus, and also represent different poetic genres.[9] Then we computed 10,000 nearest neighbours (wrt. the bigram cosine similarity metric) to every verse from that subcorpus, which gave a total of 330 million pairs from all ranges of similarity. We divided this dataset into 10 intervals based on the similarity score: $[0, 0.1), [0.1, 0.2), \ldots, [0.9, 1]$ and took random samples containing an equal number of verse pairs from each interval. Although this approach depends somewhat on the method under evaluation, it guarantees that the low-similarity ranges will also be represented, while the proportion of equivalent pairs in the sample is still sufficiently high (around 20%).

The task for the annotators was to decide whether a given pair of verses is equivalent or not. We proceeded according to typical linguistic annotation methodology (Artstein, 2017): first, the annotation guidelines were agreed in writing. Then we took a sample of 1,000 verse pairs that was annotated by each of the four annotators. In some detailed cases, deciding about the equivalence of verses turned out to be a non-trivial task and may require further discussion. However, the inter-annotator agreement (Fleiss' κ) was 0.915, which was deemed sufficient to proceed with full-scale annotation. In the second stage, each annotator received a different sample of 3,000 verse pairs, which gave a dataset of 12,000 pairs in total (the part used for assessing agreement was not used in evaluation). We have made the resulting dataset publicly available.[10]

4.2 Results

Based on the annotated sample of verse pairs, we calculated the precision and recall as, respectively, the proportion of verse pairs sharing a cluster that were truly equivalent (precision), and the proportion of equivalent verse pairs that were clustered together (recall). Table 4 shows the results of the evaluation, depending on two parameters: d—the number of dimensions (i.e. bigrams) used in the vectorization, and α—the minimum similarity for drawing an edge in the similarity graph. Each cell of the table shows the mean and standard deviation from 10 independent runs with the given parameter setting.[11] The best obtained F-scores are bolded. The results are sensitive to the threshold value α, with 0.8 being the optimal value. On the other hand, increasing the number of dimensions beyond a certain point does not improve the results, as the additional bigrams are rare and have no significant effect on the similarity values. For the optimal $\alpha = 0.8$, the difference between $d = 200$ and

$d = 300$ is within the margin of error. We also computed the results for up to $d = 500$, which were very similar, and thus are not shown.

An F-score of 81.1% is a satisfactory value, keeping in mind that evaluating one human annotator against another on the common sample yielded between 90.4% and 96.5%. Furthermore, as will be shown in the error analysis below, many false positives are still useful, while properly distinguishing them from true positives requires an in-depth knowledge of the language.

4.3 Error analysis

Tables 5 and 6 show examples of false positives and false negatives from the evaluation of the bigram+-clustering approach. The false positives in particular illustrate the difficulty of reducing the task to binary classification. Most of them are structurally similar formulas or verse types with a single word replaced or added, or common formulas consisting only of a part of a verse. Occasionally, the differing words are synonyms (*emä* ∼ *äiti* 'mother') or potentially related via derivation but with a different meaning (*ämmä* ∼ *emäntä* 'older woman'–'housekeeper'). In Table 5, only two verse pairs (pairs 10, and 13) are complete false positives, i.e. lines that do not share a relevant formulaic element. The cases of false negatives are often pairs that include significant amount of dialectal or morphological variation (*lennänä–lentee* 'fly'; *väsy–väsytän* 'get sleepy'–'I get you sleepy'), or short function words (*mill mie–joilla* 'with which (I)'; *mie voa–minä* 'I (just)'; *minkä–sen* 'what'–'that'). Similarity of content is not directly related to the similarity of bigram patterns, and both false negatives and false positives often contain cases that are highly interesting for further analysis.

Table 5. Examples of false positives

Seän ja seittemän lasta,	Tappoi seitsemän setäni lasta,
Riihimiehet jyviä mulle,	Riihimiehet mulle jyviä annoit,
Koira miulle oravan hankki,	Koira mulle orava hauku.
Vaivu maalle valkialle!	Vaivu mualle vainiolle!
Emäntä mulle kakun leipoi,	Ämmät mulle kakun leipo,
Tappo isäin, tap[po emäin],	Tappo isän, tappo äitin,
Nuku, nurmilintuseni,	Nuku, nuku, nurminukka,
Possu mulle kylkensä antoo,	Posso mulle kylkensä,
5 Minä kalat sepälle,	Minä kalat koiralle,
5 Minä hyppäsin pajaan,	10 Minä pääsin tien päälle.
Oviseinä orhin luista,	Oviseinä oravanluista,
Virolainen, vainolainen,	Venäläine vainolaine,
Nurmilintu nuuka lintu,	Nuku, nuku nurmilintu,
Väsy, vähä västäräkki,	Väsy väsy, västäräkki,

Table 6. Examples of false negatives

30 Mill' mie lennän leksuttelen,	Joilla lentee lepsuttelin
Pappi mullen paita liinan.	Pappi mulle liinapaijan,
jyvämies mnul jyvi anno,	jyvämies jyviä antoi,
Tappo isoin, tappo emoni,	Tappo issäin, tappo emmoin,
Väsy, väsy, västäräkki,	Väsytän, väsytän västäräkkii.
Silmät kieron keksinon!	Kieroin silmin keksisit,
Mie voa vierin vitsikkoo,	Minä vierin vitsikkohon,
Ken katein kahtonoo,	Ken kateen katsonee,
5 Sit sauva tiel soatto,	Sauvapa minut tielle saattoi,
Oli tappaa minunkin.	Oispa tappant minutkii,
15 Minä oravan kotkalinnul.	Minä oravan kokkolinnulle.
Nukutan, nukutan nurmi-lintuu,	Nuku, nuku, nurmilintu,
Ämmä minull' kaku leipos,	Ämmä mulle kakon leipo,
Minkä pyörä pyörähtääpi.	30 Sen pyörä pyörähtelöö,

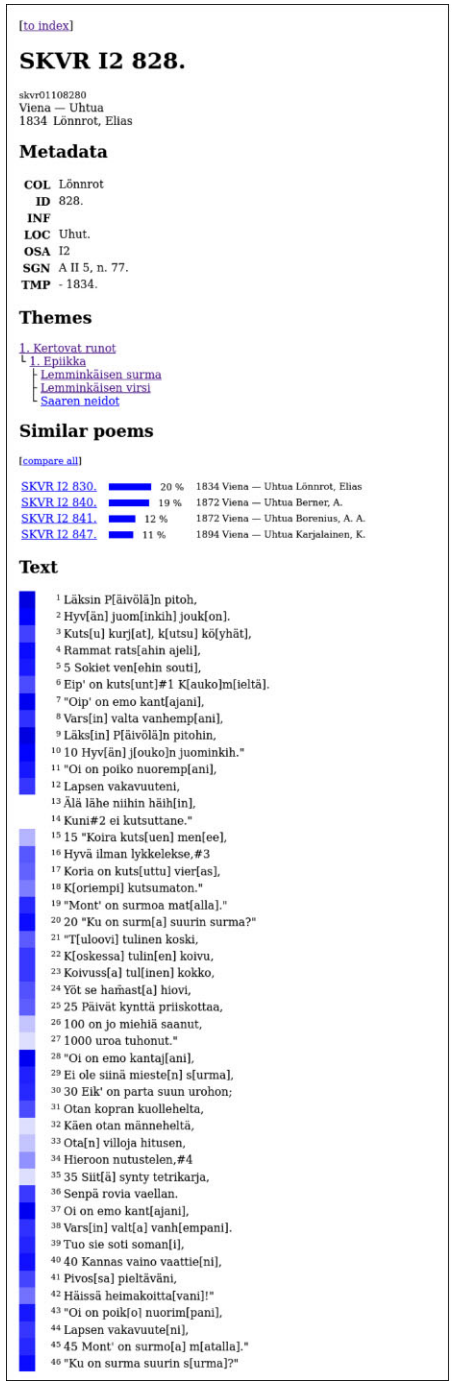
5 Application: The 'Runoregi' Browser

The precomputed similarity and clustering, together with the texts and metadata of the original corpus, are stored in a relational database and exposed via a prototype Web user interface named 'Runoregi',[12] which enables the scholar to browse through the automatically identified intertextual links. Figures 2 and 3 present the different views of Runoregi, which will be explained subsequently.

The typical entry point to the interface is the poem view (Fig. 2, left), which displays a single entire poem. In addition to the text and metadata, it includes links to automatically computed similar poems and cluster sizes for each verse.

5.1 Poem similarity and alignment

In order to compute the poem similarities, we construct a document-term matrix D, in which each poem is represented by a vector of verse cluster frequencies. Then we compute the product DD^T, which is a document-document matrix of scalar products.[13] We extract the non-zero entries of that matrix, normalize them to cosine similarities and apply a


[to index]

SKVR I2 828.

skvr01108280
Viena — Uhtua
1834 Lönnrot, Elias

Metadata

COL Lönnrot
ID 828.
INF
LOC Uhut.
OSA I2
SGN A II 5, n. 77.
TMP - 1834.

Themes

1. Kertovat runot
└ 1. Epiikka
  ├ Lemminkäisen surma
  ├ Lemminkäisen virsi
  └ Saaren neidot

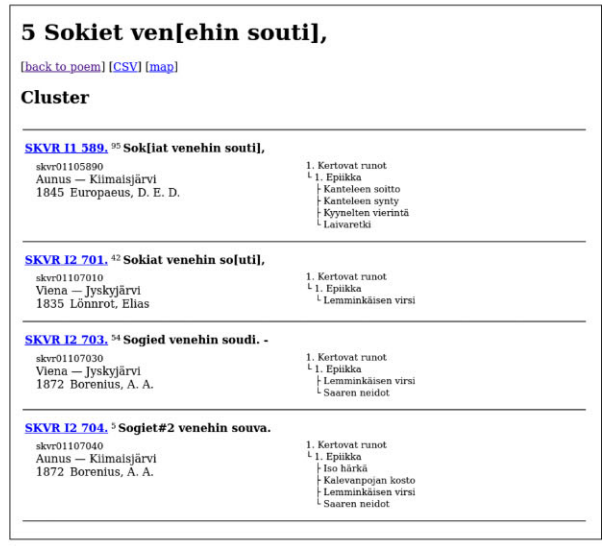
Similar poems

[compare all]

SKVR I2 830.	20 %	1834 Viena — Uhtua Lönnrot, Elias
SKVR I2 840.	19 %	1872 Viena — Uhtua Berner, A.
SKVR I2 841.	12 %	1872 Viena — Uhtua Borenius, A. A.
SKVR I2 847.	11 %	1894 Viena — Uhtua Karjalainen, K.

Text

[1] Läksin P[äivölä]n pitoh,
[2] Hyv[än] juom[inkih] jouk[on].
[3] Kuts[u] kurj[at], k[utsu] kö[yhät],
[4] Rammat rats[ahin ajeli],
[5] 5 Sokiet ven[ehin souti],
[6] Eip' on kuts[unt]#1 K[auko]m[ieltä].
[7] "Oip' on emo kant[ajani],
[8] Vars[in] valta vanhemp[ani],
[9] Läks[in] P[äivölä]n pitohin,
[10] 10 Hyv[än] j[ouko]n juominkih."
[11] "Oi on poiko nuoremp[ani],
[12] Lapsen vakavuuteni,
[13] Älä lähe niihin häih[in],
[14] Kuni#2 ei kutsuttane."
[15] 15 "Koira kuts[uen] men[ee],
[16] Hyvä ilman lykkelekse,#3
[17] Koria on kuts[uttu] vier[as],
[18] K[oriempi] kutsumaton."
[19] "Mont' on surmoa mat[alla]."
[20] 20 "Ku on surm[a] suurin surma?"
[21] "T[uloovi] tulinen koski,
[22] K[oskessa] tulin[en] koivu,
[23] Koivuss[a] tul[inen] kokko,
[24] Yöt se hamast[a] hiovi,
[25] 25 Päivät kynttä priiskottaa,
[26] 100 on jo miehiä saanut,
[27] 1000 uroa tuhonut."
[28] "Oi on emo kantaj[ani],
[29] Ei ole siinä mieste[n] s[urma],
[30] 30 Eik' on parta suun urohon;
[31] Otan kopran kuollehelta,
[32] Käen otan männeheltä,
[33] Ota[n] villoja hitusen,
[34] Hieroon nutustelen,#4
[35] 35 Siit[ä] synty tetrikarja,
[36] Senpä rovia vaellan.
[37] Oi on emo kant[ajani],
[38] Vars[in] valt[a] vanh[empani].
[39] Tuo sie soti soman[i],
[40] 40 Kannas vaino vaattie[ni],
[41] Pivos[sa] pieltäväni,
[42] Häissä heimakoitta[vani]!"
[43] "Oi on poik[o] nuorim[pani],
[44] Lapsen vakavuute[ni],
[45] 45 Mont' on surmo[a] m[atalla]."
[46] "Ku on surma suurin s[urma]?"


5 Sokiet ven[ehin souti],

[back to poem] [CSV] [map]

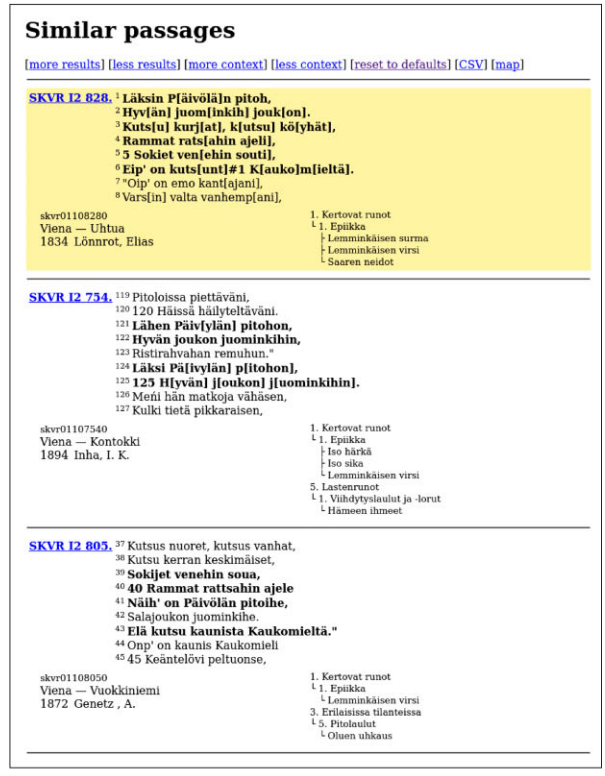
Cluster

SKVR I1 589. [95] **Sok[iat venehin souti],**
skvr01105890
Aunus — Kiimaisjärvi
1845 Europaeus, D. E. D.
1. Kertovat runot
└ 1. Epiikka
  ├ Kanteleen soitto
  ├ Kanteleen synty
  ├ Kyynelten vierintä
  └ Laivaretki

SKVR I2 701. [42] **Sokiat venehin so[uti],**
skvr01107010
Viena — Jyskyjärvi
1835 Lönnrot, Elias
1. Kertovat runot
└ 1. Epiikka
  └ Lemminkäisen virsi

SKVR I2 703. [54] **Sogied venehin soudi. -**
skvr01107030
Viena — Jyskyjärvi
1872 Borenius, A. A.
1. Kertovat runot
└ 1. Epiikka
  ├ Lemminkäisen virsi
  └ Saaren neidot

SKVR I2 704. [5] **Sogiet#2 venehin souva.**
skvr01107040
Aunus — Kiimaisjärvi
1872 Borenius, A. A.
1. Kertovat runot
└ 1. Epiikka
  ├ Iso härkä
  ├ Kalevanpojan kosto
  ├ Lemminkäisen virsi
  └ Saaren neidot


Similar passages

[more results] [less results] [more context] [less context] [reset to defaults] [CSV] [map]

SKVR I2 828. [1] **Läksin P[äivölä]n pitoh,**
[2] **Hyv[än] juom[inkih] jouk[on].**
[3] **Kuts[u] kurj[at], k[utsu] kö[yhät],**
[4] **Rammat rats[ahin ajeli],**
[5] **5 Sokiet ven[ehin souti],**
[6] **Eip' on kuts[unt]#1 K[auko]m[ieltä].**
[7] "Oip' on emo kant[ajani],
[8] Vars[in] valta vanhemp[ani],
skvr01108280
Viena — Uhtua
1834 Lönnrot, Elias
1. Kertovat runot
└ 1. Epiikka
  ├ Lemminkäisen surma
  ├ Lemminkäisen virsi
  └ Saaren neidot

SKVR I2 754. [119] Pitoloissa piettäväni,
[120] 120 Häissä häilyteltäväni.
[121] **Lähen Päiv[ylän] pitohon,**
[122] **Hyvän joukon juominkihin,**
[123] Ristirahvahan remuhun."
[124] **Läksi Pä[ivylän] p[itohon],**
[125] **125 H[yvän] j[oukon] j[uominkihin].**
[126] Meni hän matkoja vähäsen,
[127] Kulki tietä pikkaraisen,
skvr01107540
Viena — Kontokki
1894 Inha, I. K.
1. Kertovat runot
└ 1. Epiikka
  ├ Iso härkä
  ├ Iso sika
  └ Lemminkäisen virsi
5. Lastenrunot
└ 1. Viihdytyslaulut ja -lorut
  └ Hämeen ihmeet

SKVR I2 805. [37] Kutsus nuoret, kutsus vanhat,
[38] Kutsu kerran keskimäiset,
[39] **Sokijet venehin soua,**
[40] **40 Rammat rattsahin ajele**
[41] **Näih' on Päivölän pitoihe,**
[42] Salajoukon juominkihe.
[43] **Elä kutsu kaunista Kaukomieltä."**
[44] Onp' on kaunis Kaukomieli
[45] 45 Keäntelövi peltuonse,
skvr01108050
Viena — Vuokkiniemi
1872 Genetz , A.
1. Kertovat runot
└ 1. Epiikka
  └ Lemminkäisen virsi
3. Erilaisissa tilanteissa
└ 5. Pitolaulut
  └ Oluen uhkaus

Fig. 2 The different views of the Runoregi user interface. Left: the poem view; top right: the verse cluster view; bottom right: the passage view

Fig. 3 Alignment view of similar poems in Runoregi

threshold for both the raw scalar product (> 3), as well as the cosine (> 0.15). This selects candidate pairs for similar poems.

In a second step, the alignment of candidate pairs is computed using the weighted edit distance algorithm (Wagner and Fischer, 1974). The algorithm computes a maximum-weight alignment, with the weight of verse insertion/deletion being 0 and of substitution either 0 (for dissimilar verses), or ξ (for similar verses, i.e. those where $\xi > \alpha$). This algorithm produces vectors of verse IDs $\mathbf{x}, \mathbf{y}$ and a vector of weights $\boldsymbol{\xi}$, all of length m, with (x_i, y_i) being an aligned pair of verses with weight ξ_i. The *alignment similarity* is defined as the number of aligned verse pairs divided by the length of the alignment:[14]

$$s = \frac{|i : \xi_i > 0|}{m} \tag{2}$$

This yields a value between 0 and 1, which is shown in the UI as a percentage (Fig. 2, left). Pairs with similarity above 10% are stored in the database and shown in the poem view. The IDs of similar poems are links that lead to an alignment view[15] (Fig. 3) of the poem shown in the previous view with the chosen counterpart. The alignment view presents the corresponding verses of the two poems side-by-side and highlights in blue the differences between them on the character level. The non-corresponding verses are shown in grey.

Because the alignment computation for a single pair of poems is a quick operation, it is not stored in the database, but rather computed on-demand when the alignment view is rendered. Thus, it can also be displayed for a manually provided poem pair, even if it is not pre-stored as similar.

5.2 Verse clusters

The blue bar next to the poem text represents the cluster sizes for each verse, with darker shades corresponding to larger clusters (logarithmically scaled). This allows the user to easily spot verses that recur frequently across the dataset, possibly in slightly

different variants. Each segment of the bar links to a cluster page for the specific verse (Fig. 2, top-right), in which all occurrences are shown together with a short metadata section (place, year, collector, themes) and a link to the poem page where the verse occurs.

5.3 Similar passage search

Another functionality available in the poem view is passage search. If a span of verses from the poem text is selected with the mouse, another view appears, in which sequences of verses belonging to the same set of clusters as the selected sequence are presented (Fig. 2, bottom-right). The original sequence is highlighted in light yellow and the verses matching the query are shown in bold. Additionally, a configurable number of context verses before and after the sequence is retrieved.

The search is fuzzy: the verses do not have to appear in the same order, there might be other verses inserted in between, and not all clusters need to be found. These criteria are flexibly configurable by two parameters: `dist`—how far apart two matching verses can be, i.e. how many 'intruder' verses is allowed in between (default: 1), and `hitfact`—how large a proportion of the selection must be matched at least (default: 0.5). The search is implemented as follows: first, all occurrences of any cluster included in the query are retrieved from the database in the order of their occurrence in the corpus. Then, the list is filtered in a single pass to produce the passages matching search criteria (according to the parameters `dist` and `hitfact`).

The performance of the retrieval is predicated on the fact that even the largest clusters are relatively small compared to the entire corpus. Furthermore, in practice it is unlikely that verses from multiple large clusters occur next to each other. Thus, perhaps a bit counterintuitively, retrieving all potentially relevant verses is a fast operation (as the verses are indexed in the database by cluster ID) and the size of the resulting list is not problematically large. The page typically loads instantly (in less than 1 s) and relaxing the filtering criteria does not lead to longer search times, because all possible results are retrieved in any case.

5.4 Applications

Currently, Runoregi is used for several kinds of folkloristic analysis. First, it enables researchers to find variations of similar verse types—also such variations that the researcher is not able to imagine and find via word or collocation searches. Second, it looks very promising in recognizing texts that have some literary connection, such as edited versions of the same manuscript, oral texts that have served as a source material for literary publications, or literary publications that have affected the oral tradition. Such cases are extremely important to identify, not only in analysing the oral–literary relationships, but also in any further computational work, as they easily bias the results. Third, Runoregi is useful for finding and aligning different versions of the same story, especially within nearby regions, which saves a lot of time, although the alignments typically do require some manual adjusting, since typically not all the similar verses are recognized automatically. Quite naturally, across more distant regions with different languages and poetic cultures, the similarities are more difficult to find, as even similar stories are told with partly different words and verse types. Nevertheless, even just the three use cases above save considerable amounts of manual work and enable the finding of connections that would otherwise go unnoticed.

Poem similarity results can be used for reassessment of the folkloristic typology compiled by folklorists, and for automatic type detection in the case of untypologized material to be added into the database. Further, verse and song similarity indicators can be used to analyse the entire body of texts in terms of regional distribution, regional division, and to inquire into the relationship of creativity and traditionality in the oral transmission of knowledge.

The computational methods used by Runoregi could be applied also to other texts and languages. A few constraints might exist, though. As the method operates on verse level, it can be applied only to versified text. Furthermore, for long verses, the probability of two dissimilar strings sharing a large number of bigrams increases. While this was not a problem for the Finnic tetrameter, the applicability to meters involving a larger number of syllables would need to be experimentally verified. Finally, the optimal number of dimensions of the bigram vectors depends on

the size of a language's alphabet—as Finnic languages use a small set of letters, other languages could require higher-dimensional vectors. We have published the source code used for similarity computation as a Python package to facilitate further experiments.[16]

6 Conclusion

We have introduced a method for identifying equivalent verses in a corpus of oral folk poetry characterized by great variation in the texts' string representations. Cosine similarity of vectors of character bigrams could successfully identify similar pairs, which were then clustered using a Chinese Whispers algorithm to produce equivalence classes. On the basis of this similarity computation, we have built a Web interface that allows for exploration of similarities on verse, passage and poem level. We hope that this tool will facilitate the literary and folkloristic research around SKVR by providing an easy way to perform searches with good coverage despite the variation. The algorithmic underpinnings of our method are language-independent and can also be applied to other collections of versified texts with non-standardized orthography or dialectal variation.

Compared to other approaches to text reuse in Digital Humanities, our method is innovative in basing the similarity on character bigrams instead of words and taking advantage of the segmentation of poems into verses. The former feature allows us to discover similarity across linguistic variation, even where none of the words in the similar passages are identical, which frequently happens in SKVR. By basing the comparison on entire verses, we are able to efficiently process a large corpus, while the clustering of verses into equivalence classes provides us with the useful 'bag-of-verses' representation of documents.

Acknowledgements

The authors would like to thank Jukka Saarinen, Venla Sykäri, and Pihla Toivanen for their contributions to the gold standard annotation. This work was financed by the Academy of Finland research project no. 333138 'Formulaic intertextuality, thematic networks and poetic variation across regional cultures of Finnic oral poetry' and by Estonian Research Council grant no. 1288.

References

Abello, J., Broadwell, P., and Tangherlini, T. R. (2012). Computational folkloristics. *Communications of the ACM*, **55**(7): 60–70. https://doi.org/10.1145/2209249.2209267.

Artstein, R. (2017). Inter-annotator agreement. In N. Ide and J. Pustejovsky (eds), *Handbook of Linguistic Annotation*. Dordrecht: Springer, pp. 297–313.

Arukask, M. (2003). *Jutustava regilaulu aspektid: 19. sajandi lõpu setu lüroeepiliste regilaulude žanr ja struktuur (Aspects of narrative folksongs: The genre and structure of Setu lyro-epic Kalevala-metric songs from the late 19th century)*. Ph.D. thesis, University of Tartu.

Büchler, M. (2013). *Informationstechnische Aspekte des Historical Text Re-use (Information-technological aspects of historical text re-use)*. Ph.D. thesis, University of Leipzig.

Bendix, R. F. and Hasan-Rokem, G. (eds) (2012). *A Companion to Folklore*. Malden, MA: Wiley-Blackwell.

Bernstein, N., Gervais, K., and Lin, W. (2015). Comparative rates of text reuse in classical Latin hexameter poetry. *Digital Humanities Quarterly*, **9**(3).

Biemann, C. (2006). Chinese whispers - an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of TextGraphs: The First Workshop on Graph Based Methods for Natural Language Processing*.

Biemann, C. (2007). *Unsupervised and Knowledge-free Natural Language Processing in the Structure Discovery Paradigm*. Ph.D. thesis, University of Leipzig.

Bocek, T., Hunt, E., and Stiller, B. (2007). *Fast similarity search in large dictionaries*. Technical report, University of Zurich.

Broadwell, P. M., Leonard, P., and Tangherlini, T. R. (2017). 'Hvad der byggedes om dagen, blev revet ned om natten. . .': Word sequence repetition in Danish legend tradition. *Svenska Landsmål och Svenskt Folkliv (Swedish Dialects and Folk Traditions)*, **140**:9–25.

Büttcher, S., Clarke, C. L. A., and Cormack, G. V. (2010). *Information Retrieval: Implementing and Evaluating Search Engines*. The MIT Press.

Coffee, N., Koenig, J.-P., Poornima, S., Forstall, C. W., Ossewaarde, R., and Jacobson, S. L. (2013). The Tesserae project: intertextual analysis of Latin poetry. *Literary and Linguistic Computing*, **28**(2):221–28.

Foley, J. M. (1988). *The Theory of Oral Composition: History and Methodology*. Bloomington: Indiana University Press.

Frog. (2014). Degrees of well-formedness: the formula principle in the analysis of oral-poetic meters. *RMN Newsletter*, **8**:68–70.

Frog. (2016). Linguistic multiforms in Kalevalaic epic: toward a typology. *RMN Newsletter*, **11**:61–98.

Frog and Lamb, W. (eds) (2022). *Weathered Words: Formulaic Language and Verbal Art*. Harvard University Press.

Frog and Tarkka, L. (2017). Parallelism in verbal art and performance: an introduction. *Oral Tradition*, **31**(2): 203–32.

Haapoja-Mäkelä, H., Stepanova, E., and Tarkka, L. M. (2018). The Kalevala's languages: receptions, myths and ideologies. *Journal of Finnish Studies*, **21**(1 & 2):15–45.

Hakamies P. and Heimo, A. (eds) (2019). *Folkloristics in the Digital Age*. Helsinki: Academia Scientiarum Fennica.

Harvilahti, L. (1992a). *Kertovan runon keinot. Inkeriläisen runoepiikan tuottamisesta (Devices of narrative poetry: Producing Ingrian epic poetry), vol. 522 of SKST*. Helsinki: SKS.

Harvilahti, L. (1992b). The production of Finnish epic poetry – fixed wholes or creative compositions? *Oral Tradition*, **7**(1):87–101.

Harvilahti, L. (2019). History of computational folkloristics in Finland and some current perspectives. In P. Hakamies, and A. Heimo (eds), *Folkloristics in the Digital Age*. Helsinki: Academia Scientiarum Fennica, pp. 158–75.

Harvilahti, L. (2000). Variation and memory. In L. Honko (ed.), *Thick Corpus, Organic Variation and Textuality in Oral Tradition, vol. 7 of SFF*. Helsinki: SKS, pp. 57–76.

Ilyefalvi, E. (2018). The theoretical, methodological and technical issues of digital folklore databases and computational folkloristics. *Acta Ethnographica Hungarica*, **63**(1):209–58.

Jaago, T. (2016). Punane regilaulus: sõnad ja vormelid ("Red": Color term and formulaic poetic language). *Mäetagused*, **64**:9–34.

Jänicke, S. and Wrisley, D. J. (2017). Visualizing mouvance: toward a visual analysis of variant medieval text traditions. *Digital Scholarship in the Humanities*, **32**: 106–23.

Johnson, J., Douze, M., and Jégou, H. (2017). Billion-scale similarity search with GPUs. *arXiv preprint arXiv:1702.0873*.

Kallio, K., Frog, and Sarv, M. (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*, **12-13**: 139–61.

Kuusi, M. (1949). Sampo-eepos. Typologinen analyysi (The Sampo epic: A typological analysis). *Mémoires de la Société Finno-Ougrienne*, vol. XCVI. Helsinki: Suomalais-Ugrilainen Seura.

Kuusi, M. (1967). Fatalistic traits in Finnish proverbs. *Scripta Instituti Donneriani Aboensis*, **2**:89–96.

Kuusi, M. (1990). Epic cycles as the basis for the Kalevala. In L. Honko (ed.), *Religion, Myth and Folklore in the World's Epics. The Kalevala and its Predecessors*. Berlin and New York: De Gruyter, pp. 133–55.

Leino, P. (1975). Äidinkieli ja vieras kieli: Rahvaanrunouden metriikkaa (Native language and foreign language: Metrics of 19th century contemporary folk poetry). In *Mittoja, muotoja, merkityksiä*. Helsinki: SKS, pp. 207–30[2002].

Lord, A. B. (1960). *The Singer of Tales*. Cambridge: Harvard University Press. [Repr. 2000].

Meinecke, C., Wrisley, D. J., and Jänicke, S. (2019). Automated alignment of medieval text versions based on word embeddings. In *LEVIA '19: Leipzig Symposium on Visualization in Applications*.

Olsen, M., Horton, R., and Roe, G. (2011). Something borrowed: sequence alignment and the identification of similar passages in large text collections. *Digital Studies/Le Champ numérique*, **2**(1). DOI: http://doi.org/10.16995/dscn.258.

Parry, M. (1930). Studies in the epic technique of oral verse-making i: homer and homeric style. *HSCP*, **41**: 73–147.

Saarinen, J. (2018). *Runolaulun poetiikka: Säe, syntaksi ja parallelismi Arhippa Perttusen runoissa (Poetics of runo-songs: line, syntax and parallelism in the poems by Arhippa Perttunen)*. Ph.D. thesis, University of Helsinki. http://urn.fi/URN:ISBN:ISBN 978-951-51-3919-1.

Saarlo, L. (2005). *Eesti regilaulude stereotüüpiast (The stereotypy of Estonian runo-songs: Theory, method and meaning)*. Ph.D. thesis, University of Tartu. http://hdl.handle.net/10062/838.

Sarv, M. (2004). Laulikute värsivara võrdlus: metoodiline eksperiment (Comparison of singers' verse repertoire: A methodical experiment). In M. Sarv (ed.), *Regilaul – loodud või saadud?* Tartu: Eesti Kirjandusmuuseum, pp. 241–56.

Sarv, M. (2008). *Loomiseks loodud: regivärsimõõt traditsiooniprotsessis (Created for creation: Verse metre of Estonian Regilaul in the tradition process)*. Ph.D. thesis, University of Tartu.

Sarv, M. (2015). Regional variation in folkloric meter: the case of Estonian runosong. *RMN Newsletter*, **9**:6–17.

Sarv, M. (2017). Towards a typology of parallelism in Estonian poetic folklore. *Folklore: Electronic Journal of Folklore*, **67**: 65–92. https://doi.org/10.7592/FEJF2017.67.sarv.

Sarv, M. (2019). Poetic metre as a function of language: linguistic grounds for metrical variation in Estonian runosongs. *Studia Metrica et Poetica*, **6**(2):102–48.

Shang, W. and Underwood, T. (2021). Improving measures of text reuse in English poetry: A tf-idf based method. In *iConference 2021: Diversity, Divergence, Dialogue*. Switzerland AG: Springer Nature, pp. 469–77.

Shmidman, A., Koppel, M., and Porat, E. (2018). Identification of parallel passages across a large Hebrew/Aramaic corpus. *Journal of Data Mining and Digital Humanities*.

Smith D. A., Cordell, R., Dillon, E. M., Stramp, N., and Wilkerson, J. (2014). Detecting and modeling local text reuse. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 183–92.

Sturgeon, D. (2018). Unsupervised identification of text reuse in early Chinese literature. *Digital Scholarship in the Humanities*, **33**(3):670–84.

Tangherlini, T. R. (2016). Big folklore: a special issue on computational folkloristics. *Journal of American Folklore*, **129** (511):5–13. https://www.jstor.org/stable/10.5406/jamerfolk.129.511.0005.

Tarkka, L. (2013). *Songs of the Border People: Genre, Reflexivity, and Performance in Karelian Oral Poetry*. Helsinki: Suomalainen Tiedeakatemia.

Timonen, S. (2004). *Minä, tila, tunne. Näkökulmia kalevalamittaiseen kansanlyriikkaan (I, space, emotion: Kalevalametric folk lyrics)*. Helsinki: SKS.

Vesanto, A., Nivala, A., Rantala, H., Salakoski, T., Salmi, H., and Ginter, F. (2017). Applying BLAST to text reuse detection in Finnish newspapers and journals, 1771–1910. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*.

Vierthaler, P., and Gelein, M. (2019). A BLAST-based, language-agnostic text reuse algorithm with a MARKUS implementation and sequence alignment optimized for large Chinese corpora. *Journal of Cultural Analytics*, **4**(2). DOI: https://doi.org/10.22148/16.034.

Wagner, R. A. and Fischer, M. J. (1974). The string-to-string correction problem. *Journal of the ACM*, **21**(I):168–73.

Yousef, T., and Jänicke, S. (2020). A survey of text alignment visualization. *IEEE Transactions on Visualization and Computer Graphics*, **27**(2):1149–59.

Notes

1 https://skvr.fi/

2 *V* denotes the lengthening of the preceding vowel.

3 The total number of distinct bigrams in the corpus is 2202 and the frequency of the bigram with rank 300 is 22,510. It might seem as though we are removing a lot of information in this step, but the evaluation in Section 4 has shown that using more than 300 bigrams does not lead to improvement. Rare bigrams occur in a small number of verses, so they do not affect the similarity scores for the vast majority of the data. For comparison, note that standard Finnish uses only twenty alphabet letters for native words, which gives an upper bound of 400 'standard' bigrams if all combinations are possible.

4 Trigrams could be used if an additional dimensionality reduction step were applied (e.g. an autoencoder). We might explore this possibility in further work.

5 E.g. for the verses *Eipä löyä Väinämöistä* and *Peätti päiväsä mänyö*, the unigram-based representations have a cosine similarity of 0.95, mostly due to the high frequency of *ä*.

6 Cosine similarity is a widely used metric in Information Retrieval. For an introduction, see e.g. Büttcher *et al.*, 2010, section 2.2.1. The idea of using character *n*-gram-based vectorization to compute the similarity of short strings is mentioned e.g. by Bocek *et al.*, 2007, section 2.3.

7 https://github.com/facebookresearch/faiss/

8 This is one of the so-called 'small world properties'. For a review of small world properties of graphs resulting from natural language data, see (Biemann, 2007, chap. 3).

9 Neljän neidon runo 'The Song of Four Maidens' (Narrative songs), Kateen sanat 'Words of Envy' (Charms), Pakeneva 'The Fleeing One' (Songs for children), Nuku nuku nurmilintu 'Sleep, Sleep the Grassbird' (Lullabies), Kehotus laulamaan 'Exhortation to Sing' and Laulajan alkusanoja 'Beginning Words of a Singer' (Lyric songs).

10 https://github.com/hsci-r/skvr-verse-equivalence-gs

11 This is because Chinese Whispers is a randomized algorithm, so it produces a slightly different clustering on every run.

12 The interface is currently available under http://runoregi.rahtiapp.fi. The name 'Runoregi' builds on the idea of 'a sledge full of poems' common in Northern Finnic oral poetry, and the common qualifiers for the old oral

poetry in Finnic languages 'runo' and 'regi' (Finnish and Karelian: 'runolaulu' 'runo-song, song in old oral meter', 'runo' 'song, poem, singer'; Estonian:'regilaul' 'song in old traditional meter', from Low German 'Reie', 'Reige', 'dancing song').

13 The matrix D is of size $n \times c$, and DD^T is $n \times n$, with $n = 89,247$ being the number of poems and $c =$ 502,076 the number of verse clusters. Both matrices are sparse. The computation is implemented using the Python package scipy.sparse, for which those sizes do not present a problem.

14 Alternatively, the alignment similarity could be defined simply as the mean weight: $s = \frac{1}{m}\sum_{i=1}^{m} \xi_i$, which yields slightly lower values. However, we find that the similarity according to the definition given in (2) is easier to interpret.

15 For a survey of text alignment visualization methods and their applications, see Yousef and Jänicke (2020). The view presented here is classified as a 'side-by-side view' in the survey's terminology.

16 https://github.com/hsci-r/shortsim