Recognising Intertextuality in the Digital Corpus of Finnic Oral Poetry : Experiment with the Sampo Cycle

Kallio, Kati

CEUR-WS.org
2022

Kallio , K , Janicki , M M , Mäkelä , E & Sarv , M 2022 , Recognising Intertextuality in the Digital Corpus of Finnic Oral Poetry : Experiment with the Sampo Cycle . in K Berglund , M La Mela & I Zwart (eds) , Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022) . CEUR Workshop Proceedings , vol. 3232 , CEUR-WS.org , Aachen , pp. 279-287 , Digital Humanities in the Nordic and Baltic Countries Conference , Uppsala , Sweden , 15/03/2022 . < http://ceur-ws.org/Vol-3232/paper26.pdf >

http://hdl.handle.net/10138/350121

Recognising Intertextuality in the Digital Corpus of Finnic Oral Poetry: Experiment with the Sampo Cycle

Kati Kallio[1,2], Maciej Janicki[2], Eetu Mäkelä[2] and Mari Sarv[3]

[1] *Finnish Literature Society, Helsinki, Finland*
[2] *University of Helsinki, Helsinki, Finland*
[3] *Estonian Folklore Archives, Tartu, Estonia*

Abstract

While digital corpora have enabled new perspectives into the variation and continuums of human communication, they often pose problems related to implicit biases of the data and the limited reach of current methods in recognising similarity in linguistically complex data, especially in small languages.

The digital corpus of historical Finnic oral poetry in alliterative tetrametre is characterised by significant poetic, linguistic and orthographic variation. At the extreme, a word may be written in hundreds of different ways. The current corpus comprises 189,189 poetic texts in six Finnic languages (Karelian, Ingrian, Votic, Estonian, Seto and Finnish) recorded in 1564–1957 by 5,287 recorders. It has a long curation history and significant bias towards some genres, poetic forms and regions that collectors have preferred.

In this poetic tradition, an idea is typically expressed with several parallel, partly alternative poetic lines or motifs, and similar verse types may be used in different contexts. A manual attempt to find all the occurrences of widely used expressions or motifs in the corpus is an unattainable task. While the digital tools—starting from simple queries to more advanced methods—make it possible to aim at wider intertextual analyses, some part of relevant material is typically not reached. Thus, it becomes central to estimate the amount and quality of the relevant data that is not recognised with different methods.

Here, we discuss two strategies for mapping intertextuality in the corpus: 1) proceeding with text queries and 2) recognising similar poetic lines computationally, based on string similarity. We compare these approaches with one another, and then proceed to compare the results they yield with the existing type index and the results of manual early 20th-century research. While the methodological and theoretical foundations of this type of research no longer hold, and while our further interest lies in the intertextuality and variation rather than in the problematic concept of poem types, parts of earlier analyses may be used in evaluating the performance of digital approaches.

Keywords

Intertextuality, variation, similarity recognition, text reuse, oral poetry

1. Introduction

Variation is a key element of the oral tradition and human communication at large. Aiming to tackle and understand the concept when arranging and analysing the sources at a time prior to digitalisation, researchers in folklore studies have utilised several tools: notes, lists and records of genres in the manuscripts, classification of songs and stories into card files, systematized source publications, typologies and indices, various kinds of punch cards, manual tables and statistics. Despite the ardent critique of the theoretical base of earlier research that has produced systematised datasets and indices in large quantities[1,2], many of the earlier results are useful when dealing with large volumes of

The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022.
EMAIL: kati.kallio@helsinki.fi (A. 1); maciej.janicki@helsinki.fi (A. 2); eetu.makela@helsinki.fi (A. 3); mari@haldjas.folklore.ee (A. 4).
ORCID: 0000-0002-3673-1409 (A. 1); 0000-0002-8366-8414 (A. 2); 0000-0003-3981-8021(A. 3); 0000-0001-5309-2357 (A. 4).

material. Currently, researchers using the digital corpus of Finnic oral poetry either in public databases or in our Octavo interface often rely on word queries and existing type indices.

In this article, we use pre-existing type index called SKVR index and older research as a point of comparison and a way to test and develop computational methodologies to query and analyse intertextuality in the corpus of Finnic oral poetry. We concentrate on the Karelian cycle of poems called the *Sampo Cycle*, which is one of the most analysed parts of Finnic oral poetry. The parts of the cycle about the mythical object *sampo* itself are not known from Estonia or Ingria, although e.g. the forging of a golden maiden typical to the Karelian cycle was known also there [3,4].

Our aim is to analyse not only how our tools Octavo and Runoregi serve in identifying complex intertextuality, but also the principles the present SKVR index uses in identifying the *Sampo* epic, helping us to assess its usefulness. Due to the notable linguistic and poetic variation of the data of Finnic oral poetry [5–7], standard natural language processing methods are difficult to implement; thus, in this article, we restrict the discussion first to text queries and then to similarity recognition.

2. Corpus of Finnic oral poetry and the Sampo Cycle

The Finnic oral tradition, called *runosongs*, *regilaul* or Kalevalaic poetry, consists of various genres in several related languages using one poetic idiom, which has both local variations and features shared across the Finnic area. Due to historical, national and scholarly reasons, this tradition was documented extensively, mostly during the 19th and early 20th century [8].

Scholars were most interested in genres they thought were the most ancient and culturally important: heroic epics and long mythological charms. The Karelian *Sampo Cycle* was considered one of the most important among these. *Sampo* is a mythical object forged by the heavenly smith Ilmarinen. The core story consists of several parts. A typical long performance of the story in Archangel Karelia may begin with the *Creation of the World* and the *Shooting of Väinämöinen*, continue with the *Forging of the Sampo*, and the *Creation of the Kantele-instrument*, often ending with the *Theft of the Sampo*. These may aggregate in various ways with one another and with other narratives or occur separately [4,9–11]. In the SKVR index, these appear as poem types of their own (see Fig. 3 below).

Currently, the Estonian *runosong* database ERAB contains 100,034 texts [12] and the Finnish SKVR database 89,247 texts [13]; in addition, there are other digital and archival collections. SKVR is a digitized scholarly edition originally published in 1908–1948 and 1997, whereas ERAB consists of texts from the archival card files, revised according to the original manuscripts and normalized by spelling. While both corpora are rich in poetic, dialectal and morphological variation, SKVR is more diverse linguistically and orthographically. Some recurrent long words, such as Väinämöinen, may be written in over 300 different ways. The historical poetic idiom in small, related languages of our data does not have ready-made parsers or even encompassing dictionaries—the dictionaries of Estonian and Finnish dialects are in progress, of Seto and Karelian available in digital form, and of Votic and Ingrian in printed form; typically, these do not cover all the words, let alone word forms used in the poetic data. At the moment, our research database combines these two corpora, consisting of 189,189 poetic texts with more than 10 million words, exhibiting around 1 million different word forms.

In oral tradition, the amount and nature of variation itself typically varies according to tradition, region, genre, individual competence, performance situation and the intention of the performance. Variation takes place at multiple levels: a word may have been pronounced or transcribed in various ways, it may be used in various conventional collocations or verse types, and these may form various kinds of motifs that are used to build wider story patterns. Some stories and expressions may remain quite similar from performance to performance, while others are modified or used as parts of new narratives. Regularly, the collocations, motifs and story types of oral tradition do not have clear-cut borders.

3. Obtaining a corpus with text queries in Octavo

Octavo is a search interface that makes use of text and metadata queries and was developed by Eetu Mäkelä to support humanities and social science research based on combinations of large, varied and ‘noisy’ text corpora along with attendant metadata (see further [6]). In folkloristic use, its benefits are

the possibilities to use either very simple or sophisticated queries to define the amount of text and metadata shown in the query results and sort the results. It also allows researchers to choose the needed items for further examination and link the results to the original database entries but also to the Runoregi interface, which helps the researcher to identify and compare similar lines and texts across the corpus. The word *runoregi* is a metaphorical expression "a sledge for poems" in Karelian language but can also be understood as a combination of the name of this poetic genre in Finnish and Karelian (*runolaulu*) and in Estonian (*regilaul*). Both Octavo and Runoregi make use of some basic normalisation rules, helping to overcome part of the orthographic and linguistic variation. Currently, Octavo is also used among researchers outside our project (see [7,14,15]).

Gathering a type, motif or theme-based sub-corpus of Finnic oral poetry in Octavo for further analysis may include chain-like searches that make use of variants of key terms, recurring collocations (formulas), synonyms, parallel expressions and metadata. The process is highly dependent on the user's linguistic competence and prior knowledge of the characteristics of Finnic oral poetry. Using text queries, an essential issue is to evaluate their reach. Previously, we compared the analysis made with Octavo of relations of one text to a wider corpus to earlier research using the printed corpus [16,17].

In the present paper, we examine the available tools that allow a researcher (here Kallio) to identify wider intertextual networks or motifs and poetic types, starting from one central concept, '*sampo* the (mythical) object'. The search began with simple word queries and close readings of certain texts in order to identify the variation of storylines, word forms and parallel expressions relating to *sampo*, then continued to map the possible term variants used for the object.

Sampo is typically forged out of impossible ingredients by the smith Ilmarinen while proposing to the maiden of the North. The mistress of the North ties *sampo* into a stony hill, but it gets stolen. She sets after the robbers. They cause her boat run into the rocks. She makes herself a giant bird able to carry hundreds of men, takes hold of the *sampo*, and the object breaks or falls into the sea. In the data, the only recurrent parallel expression for *sampo*, often interpreted as a part of it, is *kirjokansi*, 'a multicolored/marked/ornamented cover' (see also [10, p. 204]).

The word *sampo* varies greatly, and some of its forms and variants are homonymous with other words. This means a purely computational, distant reading method using the most common variants would go astray. Searching for relevant word forms manually is a slow and tedious process that requires much close reading—proceeding from one word form to others, using more stable word beginnings (prefix queries, e.g., Octavo query: samp*) and allowing one or two of the characters of the word to change (edit distance-based queries, e.g. samppu~2).

Overall, the process meant generating and going through 1,441 different word forms and their various uses. This took three workdays, whereas making sense of the parallel expression *kirjokansi* took about one hour. This is mostly explained by three factors: 1) *sampo* is a short word taking dozens of poetic, dialectal and morphological forms (*sammoñ*, *samppuo*, *sammuo*, *šambi*, *sammen* etc.); 2) these overlap and sometimes merge with other similar sounding words (the variants of *sammas-samba*, 'a pilliar', *sammas*, 'Candidiasis or some other disease of mouth', sampi, '*strurgeon'*, *samppia-sampo-sammakko*, 'a frog', *sampsua*, 'to call', *sampsikko-sämpsikkö*, 'a species of wide hay', *sampsukka-simpsukka*, 'a pearl', *sam(m)ota*, 'to roam', *sammoin–samoin*, 'similarly', *sammua* 'to fade'); and 3) the word itself is used to mean different things. Both the original singers and the researchers have had several potential and conflicting interpretations about the *sampo*: Is it a metaphor for abundance and wealth, a magical mill, a pillar supporting the sky and the whole world, a chest full of treasures or game, a boat or something else? (See [9–11]). This debate is explained in part by the polysemic nature of the word *sampo* and the varying character of the *Sampo Cycle* in Karelian 19th century oral culture. The large field of variations and overlaps also make it impossible to make any simple word-level query that would produce only the word forms used in the sense of '*sampo* the (mythical) object'.

In contrast, *kirjokansi* (*kirjo-kanzi, kirjakañsi, kirjoikantta*, etc.) is a longer compound word varying less and not overlapping with other words, although it has various potential or explicit meanings: 'the sky', 'a book cover', 'a multicolored/ornamented cover/lid/chest /casket', 'something parallel to or describing *sampo*'. It is used in more limited contexts, but also in contexts other than those relating to *sampo*.

The final list of relevant word forms includes 53 word forms for *sampo* and 23 for *kirjokansi,* resulting in 233 texts in the data, only 161 of which actually use the terms *sampo* (144) or *kirjokansi* (76) in the sense of a (mythical) object.

4. Recognising similarities in Runoregi

Runoregi interface by Maciej Janicki [5] is designed to recognize similarity between poetic lines, passages and entire texts. It employs a representation of poetic lines as vectors of character bigrams, so that the cosine similarity of such vectors approximates the similarity of the lines in terms of content words, while being robust to some orthographic and linguistic variation. In order to recognize sets of equivalent lines, it uses the Chinese Whispers clustering algorithm [18] with the bigram-based similarities as weights. Chinese Whispers is randomized, meaning it may produce slightly different clusterings on every run.

At its present state, Runoregi offers several possibilities to explore the intertextual relationships in the corpus, starting from one poem, passage or verse. Every poem is presented with a list of similar poems, and every verse provides access to a cluster of similar verses (see [6] for a general introduction). In addition to these automatically computed similarities, the interface also allows access to poems via the Finnish SKVR and Estonian ERAB type indices and other available metadata. The bigram-based similarity metric used by Runoregi reliably identifies similar texts in linguistically and poetically close regions. However, the linguistic border between Northern and Southern Finnic languages still poses a challenge due to the differences in language, orthography and the tradition itself: similar poetic lines in Northern and Southern Finnic languages tend to contain lower amounts of shared character bigrams and the overall amount of similar lines is lower.

The functionalities of the Runoregi are useful for examining a defined set of poems, looking for close counterparts to one poem or passage or trying to understand the scale of variation one poetic line or formula may take in the wider corpus. The verse clusters typically contain poetic, morphological or orthographical variants even a competent researcher would not be able to come up with. Yet, it is not yet easy to keep a record and understand the reach of what is found in Runoregi: solving this will enable the user to retrieve sub-corpora for further analysis more quickly and extensively than with Octavo. As a first step, to give the user a wider view on the verse level variation, the recognised similarities between verses that do not end up in one cluster are now presented as ‘neighbouring clusters’ (Fig. 1, left).

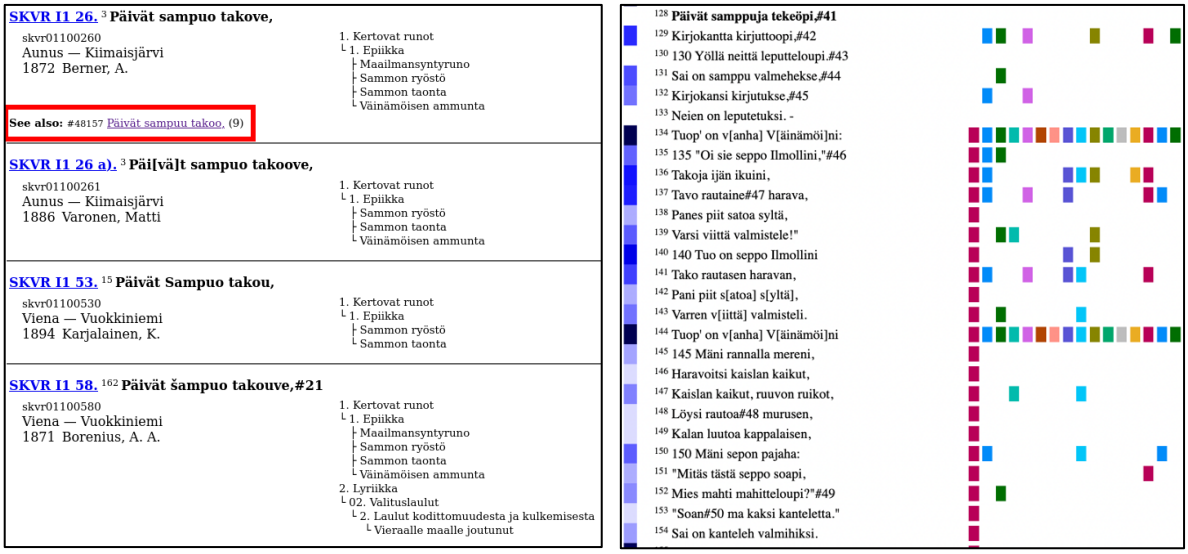

Figure 1: Runoregi views used in the present analysis Left: The verse cluster view (framed in red: a link to a neighbouring cluster). Right: The poem text with shared verses matrix view. The colour-coded cells on the right mark the occurrence of a similar verse in another poem. Poems are represented by columns and colours. The blue cells forming the bar on the left show the size of the clusters of similar verses, with darker shades representing larger sizes. Matrix view helps e.g. to understand the variation of motifs across a wide array of poems.

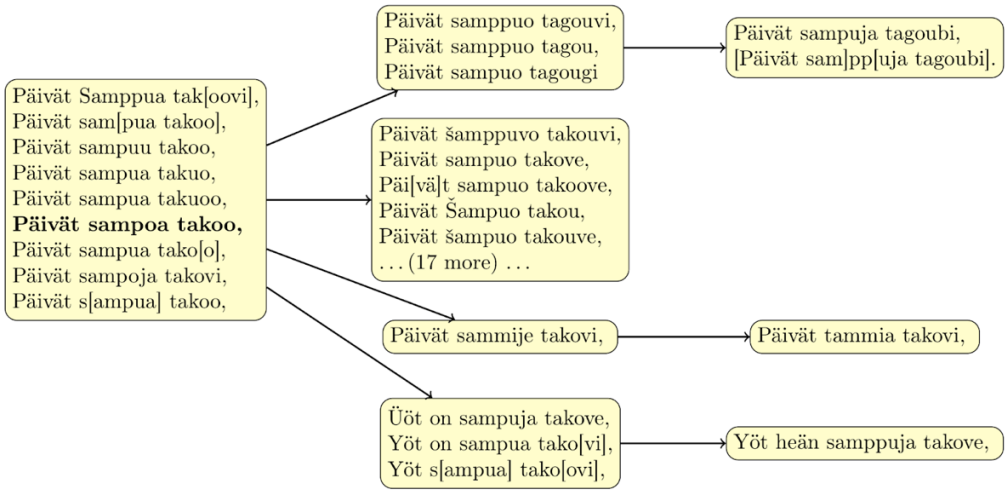

Figure 2: Browsing the network of neighbouring clusters, starting with the verse in bold leading to the leftmost cluster, then proceeding to neighbouring clusters (along the arrows) using the links provided by Runoregi.

In the *Forging of the Sampo*, one of the recurring key verses is '*päivät sampuo takovi*', or 'the days [he] is forging the *sampo'*, in different variants. The cluster including this verse form has 22 variants of the verse and links to three smaller neighbouring clusters, which provides links to some new clusters, with the final result being eight clusters and 42 verses (Fig. 2). The only line containing a word variant not obtained by the previous Octavo queries is '*Päivät tammia takovi*' or 'the days [he] is forging an oak'—this might be a mistake by the transcriber (*tammia–sammia*). In other cases, the neighbouring clusters may give other than the wanted verse types. For example, the cluster containing variants of the verse '*Kirjokantta kirjuttauvi*' or '[he] is ornamenting the ornamented cover' provides links to neighbouring clusters containing verses about taking the ornamented cover ('*Kirjokantta ottamaha*'), ornamenting an axe handle ('*Kirvesvartta kirjottavi'*), and writing in a book ('*kirjoihinsa kirjottavat*'), only the first one of these actually appearing in the *Sampo Cycle*. This means that the neighbouring clusters cannot be used in an automated way. Yet, they help a lot in finding dialectal or nearby variants.

The original idea of the present article was to try to retrieve similar sub-corpora with Octavo and Runoregi and then compare these. As it became evident that the functionalities of the Runoregi are not yet optimal for this kind of a task, the task was re-calibrated to retrieving a corpus in Octavo and analysing it with the help of Runoregi, and then comparing these processes to earlier research.

With Octavo, a single keyword and its parallel expression do not enable finding all the instances of a particular motif, let alone a theme or story—and not all the uses of them necessarily relate to the story. Indeed, part of the retrieved texts relate not to the *Sampo Cycle* proper but to *sampo* in other contexts, such as referring to the bride as a *sampo* ('wealth-producing mythical object') in wedding poems. The references using a whole line formula similar to the corresponding line type in the *Sampo Cycle* are easier to find with Runoregi than with Octavo, but keywords embedded in different line types are not. Furthermore, there are poems that represent some part or fragment of the *Sampo Cycle* without using the word for the mythical object or the key motifs. These are easier to find with Runoregi matrix view (Fig. 1, right) than with Octavo.

In checking the results obtained with Octavo in Runoregi, the biggest surprise was the number of duplicates in the data. This was identified using the 'shared verses matrix' visualisation, which marks occurrences of similar (same-cluster) verses in other texts so that wider similar passages can be identified (Fig. 1, right). At the extremes, parts of one text in the archives may have been published as six different texts that partly overlapped one another. This is due to the structure of the original book volumes organized by poem type. If a text was understood to contain several poem types, its parts were published in several places. Evidently, these duplicates may distort attempts to examine the corpus with quantitative views or map projections; thus, they need to be solved for further analyses.

Search processes with Octavo and Runoregi reveal different aspects of the *Sampo Cycle* and the corpus. Thus, using Octavo to check the results obtained with Runoregi and Runoregi to complement the Octavo queries helps improve and estimate the reach of the obtained corpus. Once the Runoregi will

include an option to more easily pick and set aside a set of relevant poems for further analysis while moving in the network of similar lines, passages and poems, it will enable gathering a sub-corpus more quickly than by word queries in Octavo.

5. SKVR index

The Octavo query-based corpus of the variants of the word *sampo* is not meant to be a complete set of texts relating to the *Sampo Cycle*. Nevertheless, testing this corpus in Runoregi and then comparing it to the present SKVR index highlights some issues relevant to challenges and possibilities in both distant and close readings of the corpus.

Any oral culture is a complex web of intertextual ties where similarity may take place at various levels, in various directions and in various quantities. There is a long and critical folkloristic discussion about the concepts of type and genre, the various potential criteria for determining these, problems with complex intertextuality, blurring of categories and the overall impossibility of creating any unambiguous or universal categorisations for wider sets of folklore (e.g. [19]). At the same time, some kind of classification and systematisation is needed, especially when dealing with large collections of texts, in order to understand the variation and fluctuations of folkloric communication in terms of centrality and marginality, on the one hand, and stability and creativity, on the other.

Both the Finnish and the Estonian corpora have type indices made along slightly different principles. In the present SKVR index, these principles also vary within the index itself. Thus, the index of narrative songs mostly follows the classification of the original book series (1908–1948), the charms are indexed along the functions and purposes of the texts rather than text types, and the lyric songs are analysed at the level of quite small motifs. The creation of the index started in the 1980s, and most of the work was done before the acquisition of the digitised data currently in use. Just having the data in digital form transforms the possibilities to understand the variation at the level of words, recurring expressions (formulas) and lines.

The Octavo word query corpus includes 161 texts, while the SKVR index gives 222 texts relating to the poem types *Forging of the Sampo* (109 texts) and the *Theft of the Sampo* (176 texts); 63 texts have both titles. The retrieved corpus included 39 texts that were not indicated with these titles in the index. In these cases, a poem may contain only a short passage mentioning *sampo* or use the concept of *sampo* in some broader proverbial or figurative sense relating to wealth or abundance (e.g. the bride or kantele-instrument referred to as *sampo*, or the bridal chest labelled as as *kirjokansi*). These occurrences help define the various conceptualisations and poetic uses of the *sampo* in 19th-century oral culture (see [11]), and even the historical layers of the concept (see [10]) but do not tell much about the story of the *sampo* itself. They are relevant, but not parts of the epic cycle about the mythical object.

The SKVR index includes 100 texts that were not captured with the word queries. Most of these are shorter fragments, not including the variants of the explicit key words. In quite many of these cases, it is not evident that the fragment should be understood as a part of the *Sampo Cycle* rather than a short charm or wedding song just referring to or reminding of the cycle. There are also some rare word variants (*sämpy* and *saima*, 3 texts) and one substitution (*tammi* ‘oak’, 1 text) of the word *sampo* that were not recognised. Finding these texts would require combing through the longer poems line by line with word queries in Octavo or, in a quicker process, using the verse matrix view and neighbouring clusters view in Runoregi.

The SKVR index is quite consistent with the most typical and explicit cases and long poems, but not so much with fragments and more hybrid texts. Irregularities seem to partly derive from the original volume indices, which were compiled by different editors using different principles for indexing. In addition, the short passages and fragments are difficult to evaluate using only the manual corpus. Indeed, Senni Timonen, who led the making of the current SKVR index, emphasised in oral communication that the index should be treated simply as a tool that helps researchers find parts of the relevant corpus for further analysis.

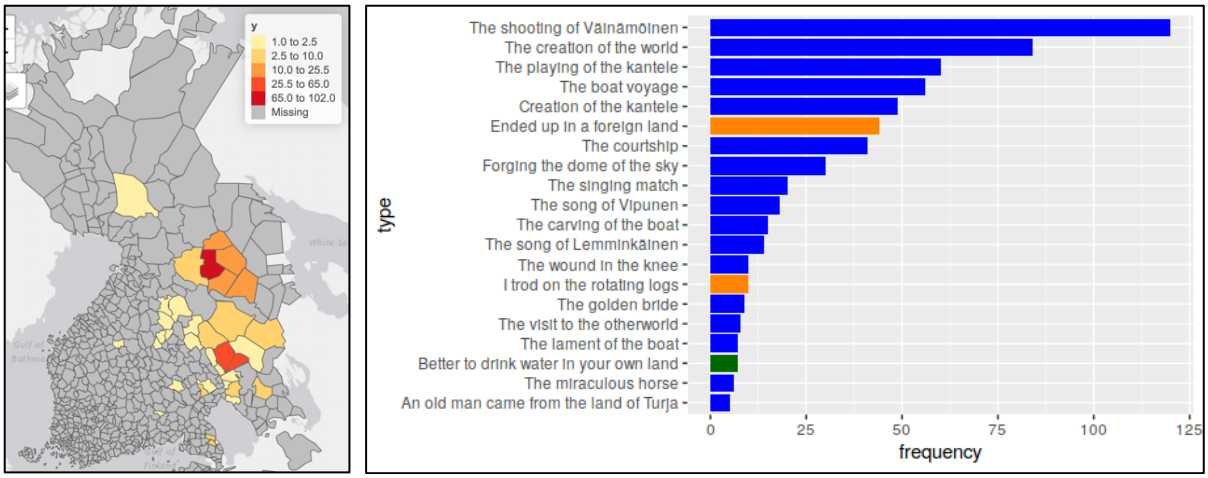

Figure 3: Distributions of the *Forging of the Sampo* and the *Theft of the Sampo* of the SKVR index. Left: The geographical distribution of these poem types. Many texts outside of Archangel and Border Karelia (around the red parishes) are not long narrative poems but fragments of parts of the *Sampo cycle* or references to it in wedding songs and charms. Right: Other poem type names of the SKVR index often appearing with the *Forging* or *Theft of the Sampo*: narrative poem types in blue, lyric ones in orange and proverbial poem types in green. The five most frequent ones and two others (*The Courtship*, *The Golden Bride*) in the list are conventionally interpreted as parts of the *Sampo Cycle*, but they also appear as separate poems and in other combinations.

6. Comparison with earlier research

The most detailed analysis of the Karelian *Sampo Cycle* and adjacent poems was published by Matti Kuusi in 1949 [4]. He read through the Karelian and Finnish parts of the SKVR series, selected a corpus of 744 poems with 41,762 verses, and conducted a complex manual analysis of recurring patterns at different levels of the cycle, also choosing a set of 6,000 verses for further analysis of poetic metre. Later scholarly analyses, even though they are conducted from very different methodological and theoretical premises, mostly rely on Kuusi's detailed descriptions and classifications of the data and concentrate on a more limited regional corpus.

The differences between the 164 poems containing variants of *sampo* or *kirjokansi*, the 222 poems noted in the SKVR index and the 744 poems used in the most comprehensible earlier study adequately demonstrate the ways both the research questions and query parameters affect the acquired corpus. Word form queries leave out fragments and some atypical variants; on the other hand, they reveal some intertextual references to the *sampo* in other poem types, and the index provides a relevant but imperfect corpus for further analysis. Kuusi includes larger body of various poems having some relation to or sharing some key motifs with the *Sampo Cycle*. These kinds of variations are easy to find with the Runoregi matrix view. Actually, a detailed analysis of the wide intertextual relationships of the varying motifs, verse types and formulas used in the *Sampo Cycle* would clearly exceed the material analysed by Kuusi. At the time of manual research, it was not possible to find all the instances of a particular verse type in the corpus of this size, and the researchers did not deem common overall formulas interesting. For example, the variants of the recurrent verse type in the *Sampo Cycle* '*Vaka vanha Väinämöinen*' or 'the steady old Väinämöinen" form the biggest verse cluster in Runoregi with 4558 verses, casting an intertextual network across poem types in various genres.

Trying to form a more complete corpus of the whole *Sampo Cycle* with Octavo or Runoregi, the next step would be to analyse which motifs and themes in the retrieved poems closely relate to the story of *sampo* itself (instead of just aggregating with it) and search for similar elements in the corpus. Yet, due to the complex and extremely varying and aggregating character of the *Sampo Cycle* as a whole, this kind of analysis gets quite interpretive.

The Karelian singers themselves did not have a concept of the *Sampo Cycle* similar to what has been used by the researchers and in the SKVR index but they had varying local and situational conventions of combining stories and motifs into performances, and habits of referring to other stories and genres

while singing a song or reciting a charm. Indeed, the present test also demonstrates how subjective and research-specific it is to define the *Sampo Cycle*. What motifs should be understood as central to the cycle, and when should short text fragments be interpreted as belonging versus referring to the cycle?

7. Discussion

Examining and comparing the present tools Octavo and Runoregi demonstrates the benefits of having several methods to access the data of Finnic oral poetry, each tool producing a different, complimentary view that helps to assess the coverage of the results.

Retrieving a corpus related to a particular concept (e.g. forest, see [15]) is something researchers of Finnic oral poetry often need. This approach is made easier by digital interfaces but requires time for cross checking and close reading the data. Due to the variation of word forms and motifs and the use of synonyms, metaphors, metonyms and other poetic expressions, the chosen queries may only cover a fraction of the potential data, and it is difficult to estimate whether all the significant sections of the data are reached. On the other hand, homonymous word forms often add irrelevant results to the results of word queries.

The text queries with the variants of the key terms (Octavo) only cover the most evident part of the intertextual network relating to some key motifs of the *Sampo Cycle*. Finding relevant but more fragmentary or partial poems requires identifying a wider set of recurrent motifs, verse types and collocations typical to the cycle. Depending on the character of the terms used in the process, word form variation and polysemy may mean that obtaining and interpreting the results require a significant amount of close reading.

The verse similarity recognition based on algorithmic clustering of bigram vectors (Runoregi) often reveals word forms, substitutions and parallel expressions that are not easily found in direct text queries. The rough and quick identification of similar texts and passages helps in understanding how the parts of a chosen text may appear in other poetic contexts. These methods also help identify problems and biases in the data, such as unexpected duplicates. Yet, at the present state of the Runoregi interface, it is difficult to quickly analyse wider, versatile subject areas such as the *Sampo Cycle*, or proceed from line or text to another in order to gather a large sub-corpus.

Earlier research and indices give a view on how some motifs and types have been interpreted in manual analysis based on close reading and manual analysis. These may help in gathering a relevant corpus but need careful analysis rather than being just used as an easy basis for new quantitative views or visualizations.

8. Acknowledgements

The research was funded by the Academy of Finland projects 333138 and 346342 and the Estonian Research Council project (PRG1288).

9. References

[1] R. D. Abrahams, Phantoms of Romantic Nationalism in Folkloristics, Journal of American Folklore 106 (1993): 3–37. https://doi.org/10.2307/541344.

[2] U. Wolf-Knuts, P. Hakamies, The intellectual and social history of folkloristics, ethnology and anthropology in Finland, in: A. Barrera-González, M. Heintz, A. Horolets (Eds.), European Anthropologies, Berghahn Books, New York, 2017, pp. 149–168. https://doi.org/10.2307/j.ctvw04gmt.11.

[3] K. Kouvola, The artificial bride on both sides of the Gulf of Finland: The Golden Maiden in Finno-Karelian and Estonian folk poetry, in: M. Bertell, Mr. Frog, K. Wilson (Eds.), Contacts and Networks in the Baltic Sea region, Amsterdam University Press, Amsterdam, 2018, pp. 211–234. https://doi.org/10.1515/9789048532674-014.

[4] M. Kuusi, Sampo-eepos: typologinen analyysi, Suomalais-ugrilainen Seura, Helsinki, 1949.

[5] M. Janicki, K. Kallio, M. Sarv, Exploring Finnic written oral folk poetry through string

similarity, Digital Scholarship in the Humanities, (in press).
[6] K. Kallio, E. Mäkelä, M.M. Janicki, Historical Oral Poems and Digital Humanities: Starting with a Finnish Corpus, Folklore Fellows Network 54 (2020): 12–18. https://www.folklorefellows.fi/historical-oral-poems-and-digital-humanities/.
[7] V. Sykäri, Digital Humanities and How to Read the Kalevala as a Thematic Anthology of Oral Poetry, Arv. (2020): 29–54.
[8] K. Kallio, M. Frog, M. Sarv, 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 (2017): 94–117. http://hdl.handle.net/10138/305420.
[9] S. Apo, Samporunojen satuainekset, in: N. Hämäläinen, P. Kauppi (Eds.), Paradigma. Näkökulmia tieteen periaatteisiin ja käsityksiin, Suomalaisen Kirjallisuuden Seura, Helsinki, 2021, pp. 89–112. https://doi.org/10.21435/ksvk.100.
[10] Mr. Frog, Confluence, continuity and change in the evolution of mythology: the case of the Finno-Karelian Sampo-cycle, in: Mr. Frog, A.-L. Siikala, E. Stepanova (Eds.), Mythic discourses. Studies in Uralic traditions, Suomalaisen Kirjallisuuden Seura, Helsinki, 2012, pp. 205–254. https://doi.org/10.21435/sff.20.
[11] L. Tarkka, The Sampo: myth and vernacular imagination, in: Frog, A.-L. Siikala, E. Stepanova (Eds.), Mythic discourses. Studies in Uralic traditions, Suomalaisen Kirjallisuuden Seura, Helsinki, 2012, pp. 143–170. https://doi.org/10.21435/sff.20.
[12] J. Oras, M. Sarv, L. Saarlo, ERAB. Eesti regilaulude andmebaas, 2003–. https://www.folklore.ee/regilaul (accessed February 14, 2022).
[13] J. Saaarinen, A. Krikman, SKVR database, Suomalaisen Kirjallisuuden Seura, Helsinki, 2004. https://skvr.fi/.
[14] N. Hämäläinen, R. Holopainen, M. Luhtala, J. Saarelainen, V. Sykäri, eds., Avoin Kalevala, Suomalaisen Kirjallisuuden Seura, Helsinki, 2019. https://kalevala.finlit.fi/.
[15] T. Seppä, Katsooko metsä ihmistä? Suomen Kansan Vanhat Runot ja lajienvälinen kommunikaatio, Kirjallisuudentutkimuksen Aikakauslehti Avain, 17.4 (2020): 22–45. https://doi.org/10.30665/av.98306.
[16] K. Kallio, E. Mäkelä, Suullisen runon sähköisestä lukemisesta, Elore 26.2 (2019): 26–41. https://doi.org/10.30666/elore.84570.
[17] K. Kallio, Oisko linnut lentoneuot: Kantelettaren runo ja taustavaikutteiden verkosto, in: U. Piela, P. Hakamies, P. Hako (Eds.), Eurooppa, Suomi, Kalevala. Mikä mahdollisti Kalevalan?, Suomalaisen Kirjallisuuden Seura, Helsinki, 2019, pp. 175–199. http://hdl.handle.net/10138/332065.
[18] C. Biemann, 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, Association for Computational Linguistics, New York City, 2006: pp. 73–80. https://aclanthology.org/W06-3812.
[19] R. Bauman, A world of others' words: cross-cultural perspectives on intertextuality, Blackwell Publishing, Oxford, 2004. DOI:10.1002/9780470773895.