Text similarity in oral runosong tradition

Towards a large-scale quantitative analysis

Maciej Janicki
Dep. of Digital Humanities
University of Helsinki

Mari Sarv
Estonian Folklore Archives
Estonian Literary Museum

Estonian Runosongs Database[1]

- songs from oral tradition in Finnic tetrameter
- 99,938 texts, 1,993,426 lines
- hierarchical type index (ongoing work)
- dialectal variation

Also in our project (but not considered here):

- similar collections from Finland (*Suomen Kansan Vanhat Runot*)
- unpublished archival materials from Finland
- literary works (e.g. Kalevala, Kalevipoeg)

Mõistke, mõistke, mehed noored
Teadke, teadke naesed targad
Arvake küla õeksed
Mis see seal mäella kasvab
Tamme seal mäella kasvab
Mis sest tammest tehtaneksi:
Tüvikust saab tünderida
Otsast õlle poolikuida
Keskelt saab kerstulaudu
Ladvast lapse kätki laudu
Mis jäevad jätiksed järele
Veeme soosse, teeme sauna
Kus need virved vihtilevad
Aned luida audunevad
Peasukesed päid pesevad
Lõukesed loputavad.

EÜS VII 313 (67)

[1] https://www.folklore.ee/regilaul/andmebaas/

Intertextuality in oral tradition

Main thesis of **oral-formulaic theory** by Milman Parry and Albert Lord:

a ‘text’ (of an oral poem) does not exist as a stable entity, instead it is recreated on every performance.

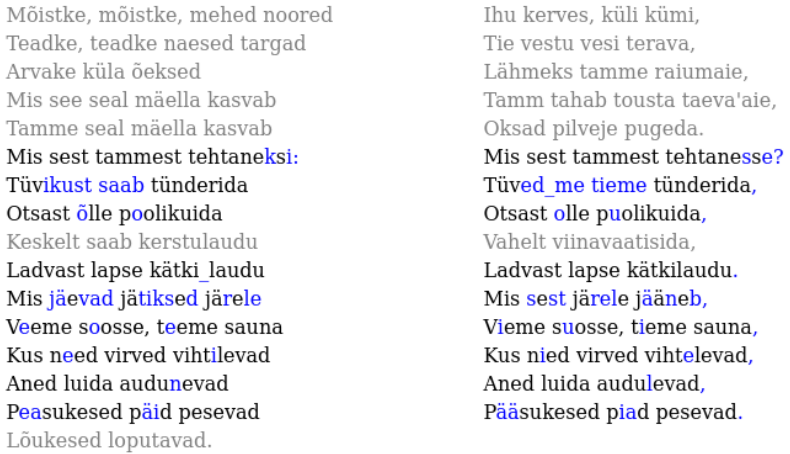
Widely known lines, formulas and motifs appear in many texts, possibly stringed together in different ways.

If this is the case, then we expect:

- very frequent partial and non-exact similarities between texts,
- but rarely complete duplicates.

Automatic alignment

- Line similarity metric: cosine similarity of character bigram vectors[1]
- Pairwise sequence alignment: Wagner-Fischer algorithm (a.k.a. weighted edit distance)[2]



[1] M. Janicki, K. Kallio, M. Sarv. Exploring Finnic written oral folk poetry through string similarity. *Digital Scholarship in the Humanities*, 2022. https://doi.org/10.1093/llc/fqac034

[2] M. Janicki. Optimizing the weighted sequence alignment algorithm for large-scale text similarity computation. (To appear in:) *Proceedings of the 2nd Workshop on Natural Language Processing for Digital Humanities (NLP4DH)*, 2022.

Similarity metrics

Mõistke, mõistke, mehed noored	Ihu kerves, küli kümi,	0
Teadke, teadke naesed targad	Tie vestu vesi terava,	0
Arvake küla õeksed	Lähmeks tamme raiumaie,	0
Mis see seal mäella kasvab	Tamm tahab tousta taeva'aie,	0
Tamme seal mäella kasvab	Oksad pilveje pugeda.	0
Mis sest tammest tehtaneksi:	Mis sest tammest tehtanesse?	0.93
Tüvikust saab tünderida	Tüved me tieme tünderida,	0.51
Otsast õlle poolikuida	Otsast olle puolikuida,	0.82
Keskelt saab kerstulaudu	Vahelt viinavaatisida,	0
Ladvast lapse kätki laudu	Ladvast lapse kätkilaudu.	0.95
Mis jäevad jätiksed järele	Mis sest järele jääneb,	0.64
Veeme soosse, teeme sauna	Vieme suosse, tieme sauna,	0.78
Kus need virved vihtilevad	Kus nied virved vihtelevad,	0.88
Aned luida audunevad	Aned luida audulevad,	0.90
Peasukesed päid pesevad	Pääsukesed piad pesevad.	0.84
Lõukesed loputavad.		0

$sim_{raw} = 7.26$

$sim_{left} = 7.26 / 16 = 0.454$

$sim_{right} = 7.26 / 15 = 0.484$

$sim_{symmetric} = (2 * 7.26) / (16+15) = 0.468$

Criteria to be considered “similar”:

- $sim_{raw} > 2$
- $sim_{left} > 0.1$ OR $sim_{right} > 0.1$

Quantitative analysis

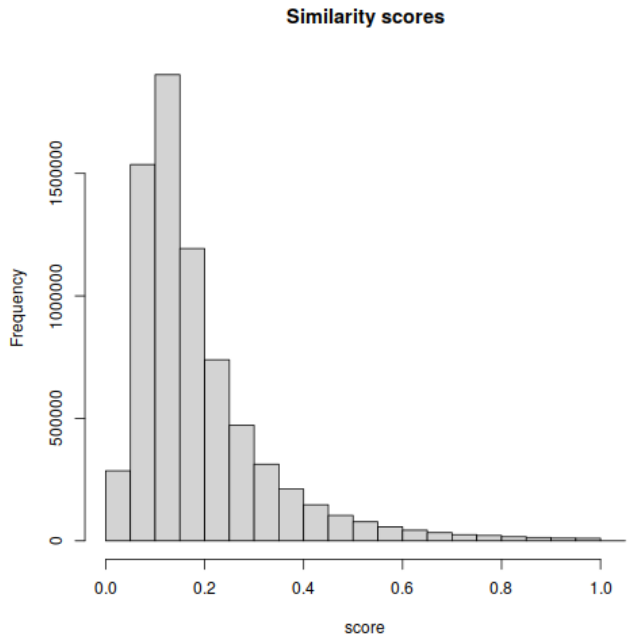
In total – 7.2 million poem pairs fulfill the criteria. Here we study:

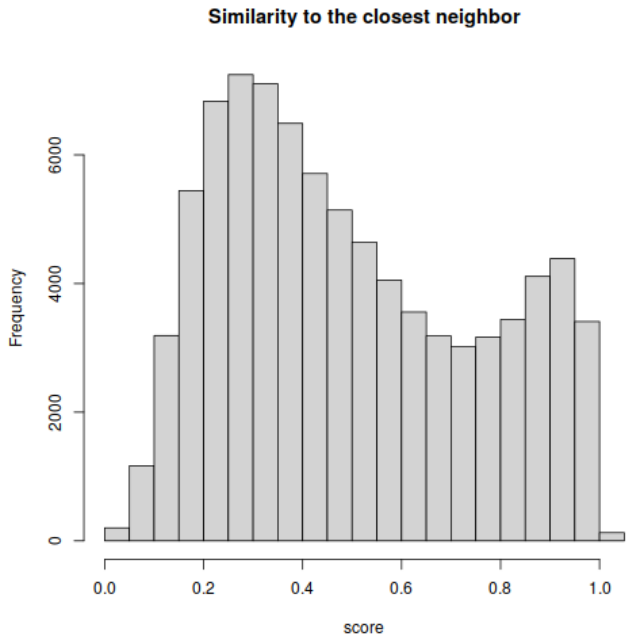
1. The continuum of similarity
   identical – similar – partially similar – short common passages

2. Local structure
   How many neighbors do poems typically have?
   How is the similarity of the neighbors distributed?

3. Comparison to type index
   Are similar texts indexed similarly? -> in general, no

The continuum of similarity





The continuum of similarity

Oleks minu olemine
Teiseks minu tegemine
Küll ma teaksin, mis ma teeksin
Ma teeks talli taeva alla
Talli sisse lateri_ja
Laterisse laugu täku
Laugu täkule sadula
Sadulasse saksa poisi
Saksa poisile kübara
Kübaralle kuldapaela.

96%

Oleks minu olemine
teiseks minu tegemine
küll ma teaksin mis ma teeksin
ma teeks talli taeva alla
talli sesse lateri
ja_laterisse laugu täku
laugu täkule sadula
sadulale saksa poisi
saksa poisile kübara
kübarale kuldapaela.

Oleks minu olemine
Teiseks minu tegemine
Küll ma teaksin, mis ma teeksin
Ma teeks talli taeva alla
Talli sisse lateri ja
Laterisse laugu täku
Laugu täkule sadula
Sadulasse saksa poisi
Saksa poisile kübara
Kübaralle kuldapaela.

37%

Laugu täkule sadula
Sadulasse saksa meesi
Saksa meesile kübara
Kübarale kuldakrooni
Kuldakroonile kuduja
Kudujale terav mõeka
Kuldamõegale tegija
Tegijale terav kerves

Oleks minu olemine
Teiseks minu tegemine
Küll_ma teaksin,_mis_ma_teeksin

Ma teeks talli taeva alla
Talli sisse lateri_ja
Laterisse laugu täku
Laugu täkule sadula
Sadulasse saksa poisi
Saksa poisile kübara
Kübaralle kuldapaela.

63%

Oleks see_mu olemine
Teiseks minu tegemine
Ma teeks_ilma_ümmarguseks,
Taevariigi ruuduliseks,
Ja teeks talli taeva alla,
Talli sisse latterad,
Latterisse laugud ruunad,
Laugud ruunad sadulas.
Sadulatel saksa poisid,
Saksa poistel kübarad,
Kübaratel kullast_poordid,
Kuldse poordil kudujad
Kudujatel kuued teljed
Kanguritel kaheksad teljed

Oleks minu olemine
Teiseks minu tegemine
Küll ma teaksin, mis ma teeksin
Ma teeks talli taeva alla
Talli sisse lateri_ja
Laterisse laugu täku

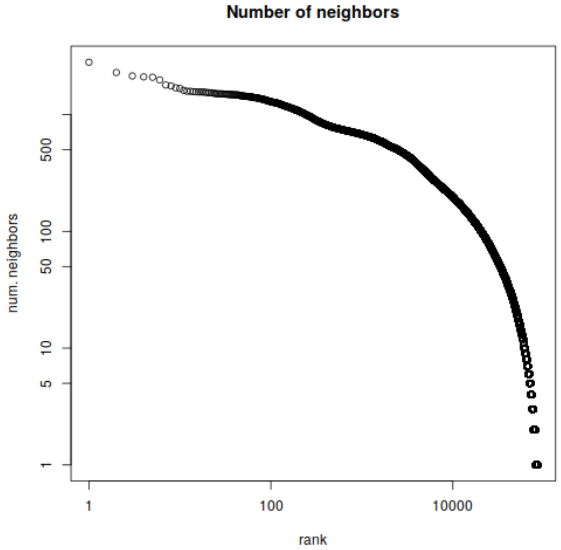
15%

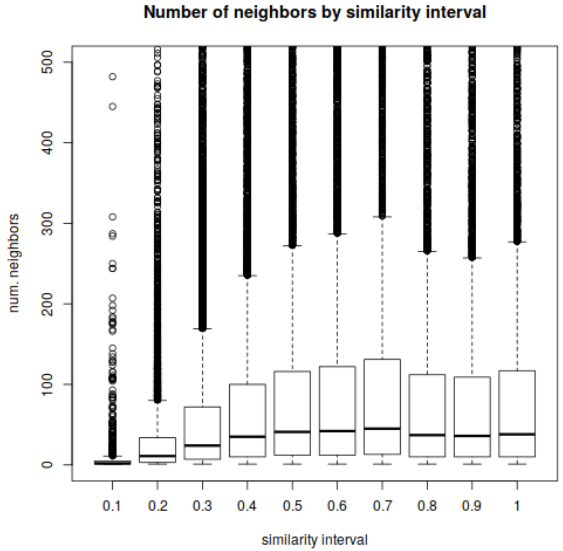
Laugu täkule sadula
Sadulasse saksa poisi
Saksa poisile kübara
Kübaralle kuldapaela.

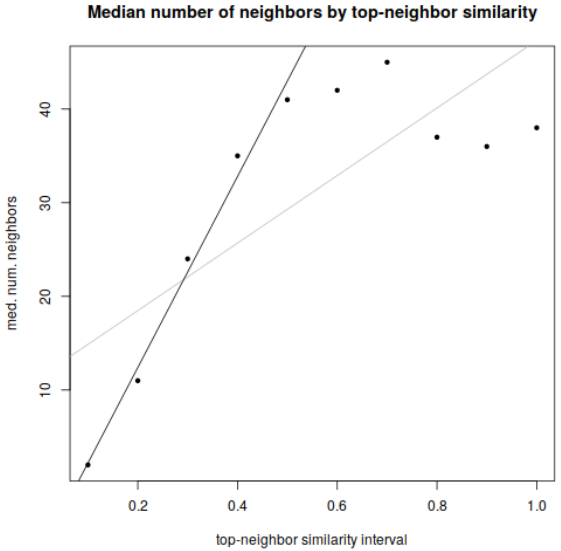
Oh seda venda mis minule
Ja seda ratsu, mis minu ver
Vend oli isi ilmatarka,
Ta tegi talli taeva alla
Talli sisse lateruse,
Laterusse lauku täkku.
Ilm ei jõudnud ehitada,
Vald ei jõudnud vallid teha,
Kihelkond kinni pidada.
Mina vennal üttelema:
Vennakene, ellakene,
Vala tal pähä vaskipäitsed,
Pane tal luine looka peale
Pane kaela raudarangid,
Mine siis soosse sõitema
Arusse teeda ajama.
Kui põle soosta sõidetud
Ja arusta teed aetud:
Mine siis läbi alta ilma,
Alta ilma, pealta päeva,
Viie vikerkaare vahelt
Ja kuie kuu keske'elt!

Local structure

- Number of neighbors: very skewed distribution (but NOT a power law)
- Poems with a highly similar neighbor tend to have more neighbors overall? (BUT: trend only on medians!)

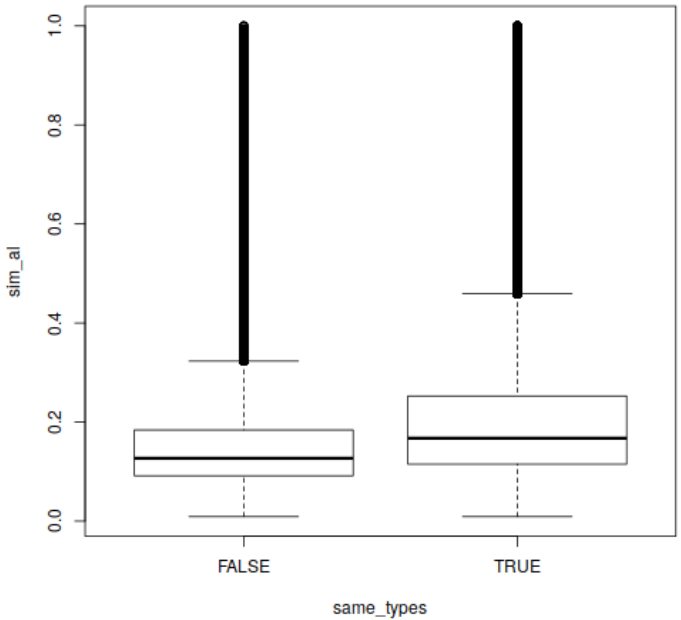






Comparison to type index

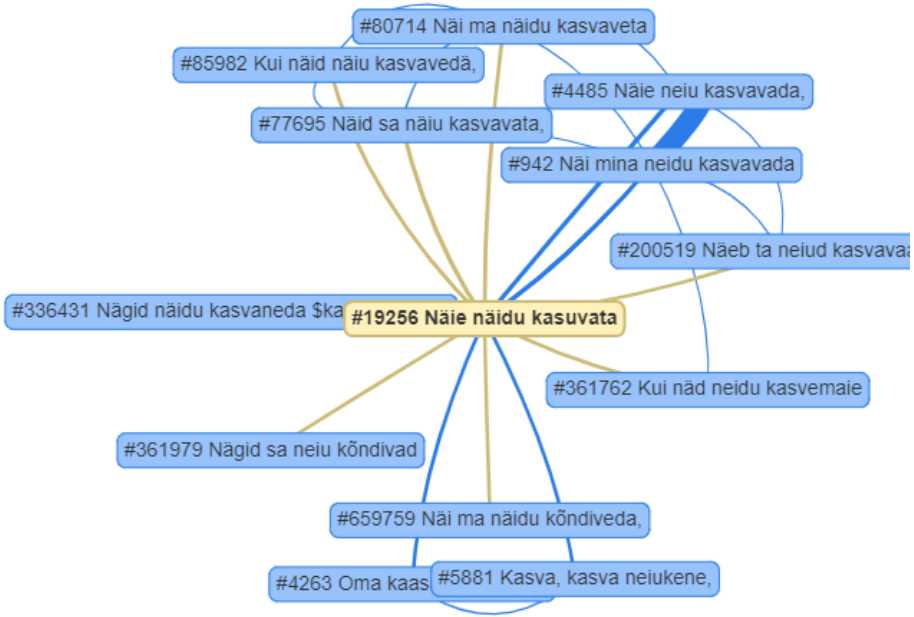
- Poems indexed in the same way are on average only slightly more similar than those indexed differently
  -> typically high textual differences also within types
- Occasionally, very similar texts are indexed differently
- Around 30% poems are missing type annotations
- Conclusion: similarity computations provide *different information* than the type index

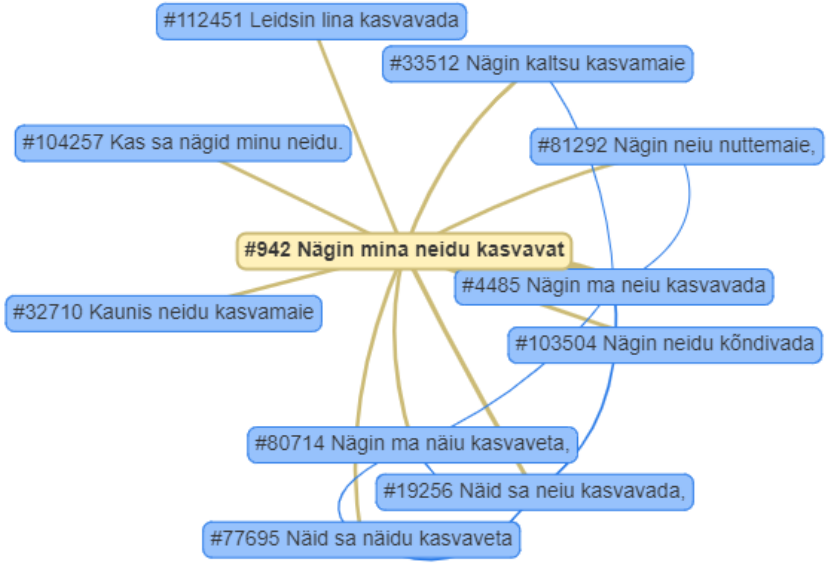


Reasons for differences

- folkloristic song types are not hermetic units, they
  - often have overlapping parts, from small formulaic units that occur in many different song types up to the sister-songtypes that have similar plot but different end-solution
  - often blend many song types into more or less coherent wholes
- due to the dialectal variation the automatic recognition of similar lines is always not possible

“I saw a young maiden growing up ...” (South/North Estonian)





Näie näidu kasuvata

Nägin mina

Conclusions

1. Automatic text alignment provides new and useful information.

2. A quantitative overview of automatically computed text similarities confirms what has been known in folkloristics:
   - fluidity of texts,
   - frequent reuse (but also variation) of shorter motifs and formulas.

3. A more detailed quantitative analysis is possible:
   - text similarities in relation to various metadata (place/singer/collector etc.)
   - inspecting outliers based on various criteria
   - network analysis
   - …

See more or contact us:
https://blogs.helsinki.fi/filter-project/

‘Formulaic intertextuality, thematic networks and poetic variation across regional cultures of Finnic oral poetry’

Photo: J. Lukkarinen, Finnish Heritage Agency