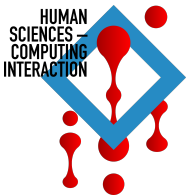
Text similarity and alignment in the study of Finnic oral folk poetry

Maciej Janicki

July 6, 2022



HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

THE FILTER PROJECT

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

A cooperation of:

1. **Finnish Literature Society**
   Kati Kallio, Jukka Saarinen
2. **University of Helsinki, HSCI group**
   Eetu Mäkelä, Maciej Janicki
3. **Estonian Literary Museum**
   Mari Sarv, Liina Saarlo

THE CORPUS

Text collections:

1. Suomen Kansan Vanhat Runot (SKVR)
2. Eesti Regilaulude Andmebaas (ERAB)
3. SKS's corpus of unpublished poems (JR)
4. Printed works: Kalevala, Kanteletar, Kalevipoeg

Characteristics:

- written in Finnic tetrameter
- mostly 19th-early 20th century
- languages: Finnish, Estonian, Karelian, Ingrian, Votic, Seto, ...

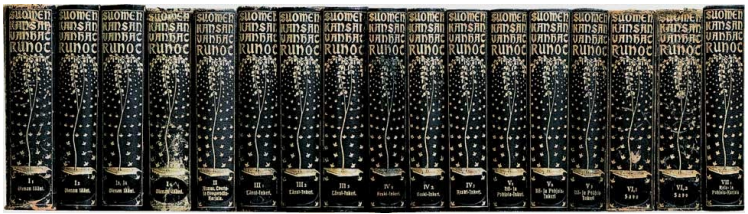
Lappalaiṅe on kyttösilmä
Piti viikoista vihuva,
Kauvon aijasta katsetta.
Huol'itti tulista nuolda
Tul'izilla jändevillä,
Peäškön pieńillä sulilla,
Varposen vivuttsimilla.
SKVR I1 17

Mõistke, mõistke, mehed noored
Teadke, teadke naesed targad
Arvake küla õeksed
Mis see seal mäella kasvab
Tamme seal mäella kasvab
EÜS VII 313 (67)

SIZE

Currently (SKVR + ERAB + JR + Kalevala + Kanteletar + Kalevipoeg):

- 275,181 documents (poems),
- 4,369,489 verses (tokens),
- 3,477,074 unique verses.



(– 15% of our corpus)

SIZE

Currently (SKVR + ERAB + JR + Kalevala + Kanteletar + Kalevipoeg):

- 275,181 documents (poems),
- 4,369,489 verses (tokens),
- 3,477,074 unique verses.

GOAL: ALIGNMENT

Yht' ei kuttsun Lemmingäistä.	15 Kuttšu veri-sogeat,
*Rujot (ne) reillä reissuaabi,	Ruiot re'ellä rembutteli,
Rammat rattšahin ajeli,	Rammat rattšahin ajeli,
20_Sogiat venozin soudi.*	Sogeat venosin souti,
	Yht' ei kuttšu Lemmingästä.
Lemmingäin on poiga_piilo	20_Lemmingäne on piilopoiga
Pillojah on piilemässä,#10	Pilloja on piilemässä,
Pahojah pagenemassa.#11	Pahoja pagenemassa.
"Hoib om moamo, kandajańi,	"Hoi on moammoni, kantajani,
25_Armas maijon andajańi,	Armas maion antajani,
Ihalan imettäjäni,	25_Ihala imettäjäni,
Ettsib om miul pelvoi paid[a],	
Ennembä neidona kuvottu,	
Kassabapeän#12 on kalkuteltu,	Tuos miull' sot'isobani,
30_Kannabas paloni_paid[a]."	Kannas paloini-paita!"

http://runoregi.rahtiapp.fi/poemdiff?nro1=skvr07108030&nro2=skvr07108080

1. align similar texts line-by-line
2. compute similarity measures summarizing the alignment
3. automatically find pairs of similar texts in the corpus
4. get large-scale overviews of the identified similarities

METHODS

1. **Verse similarity: bigram-based embedding + cosine similarity**
2. Poem similarity: weighted edit distance and alignment

VERSE SIMILARITY

Bag-of-bigrams representation:

	ei	ip	ib	pä	bä	ä_	lö	öy	öu	so	yä	uv	op	
eipä_löyä_väinämöistä	1	1	0	1	0	2	1	1	0	0	1	0	0	...
eibä_löuvä_väinämöistä	1	0	1	0	1	2	1	0	1	0	0	1	0	
eipä_sopi_väinämöistä	1	1	0	1	0	1	0	0	0	1	0	0	1	

- using d most frequent bigrams (e.g. $d = 300$)
- insensitive to word order!

VERSE SIMILARITY

Cosine similarity:

		ei	ip	ib	pä	bä	ä_	lö	öy	öu	so	yä	uv	op	
x	eipä_löyä_väinämöistä	1	1	0	1	0	2	1	1	0	0	1	0	0	
y	eibä_löuvä_väinämöistä	1	0	1	0	1	2	1	0	1	0	0	1	0	...

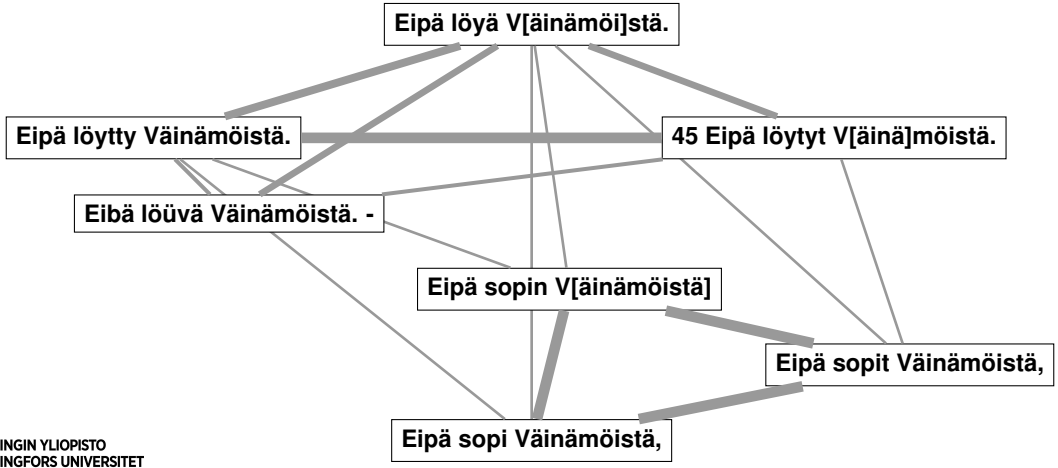
$$\begin{aligned}\cos\angle(\mathbf{x},\mathbf{y}) &= \frac{\mathbf{x}\cdot\mathbf{y}}{|\mathbf{x}|\cdot|\mathbf{y}|} = \frac{\sum_{i=1}^{d} x_i y_i}{\sqrt{\sum_{i=1}^{d} x_i^2}\sqrt{\sum_{i=1}^{d} y_i^2}} \\ &= \frac{1\cdot 1 + 1\cdot 0 + 0\cdot 1 + 1\cdot 0 + 0\cdot 1 + 2\cdot 2 + \ldots}{\sqrt{1^2+1^2+0^2+1^2+0^2+2^2+\ldots}\cdot\sqrt{1^2+0^2+1^2+0^2+2^2+\ldots}} \\ &\approx 0.81\end{aligned}$$

VERSE SIMILARITY

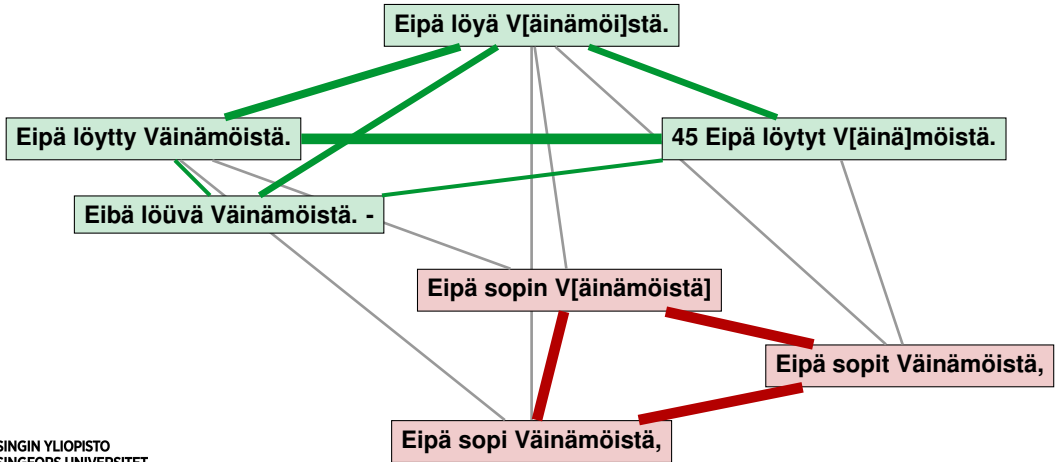
Query: Eipä löyä V[äinämöi]stä
Result:

Eipä löytty Väinämöistä.	0.863636
45 Eipä löytyt V[äinä]möistä.	0.826869
Eibä löuvä Väinämöistä. -	0.826869
Eipä löütüt Väinämöistä.	0.800197
Eipä sopi Väinämöistä,	0.76277
Eipä sopin V[äinämöistä]	0.755742

CLUSTERING

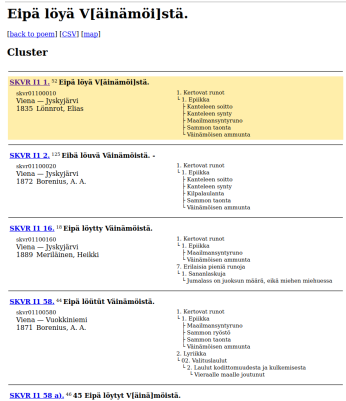

Eipä löyä V[äinämöi]stä.
Eipä löytty Väinämöistä.
45 Eipä löytyt V[äinä]möistä.
Eibä löüvä Väinämöistä. -
Eipä sopin V[äinämöistä]
Eipä sopit Väinämöistä,
Eipä sopi Väinämöistä,

CLUSTERING


Eipä löyä V[äinämöi]stä.
Eipä löytty Väinämöistä.
45 Eipä löytyt V[äinä]möistä.
Eibä löüvä Väinämöistä. -
Eipä sopin V[äinämöistä]
Eipä sopit Väinämöistä,
Eipä sopi Väinämöistä,

CLUSTERING[1]

51 Talvet lylyn livulla;
52 **Eipä löyä V[äinämöi]stä.**
[illegible]ni seppol[an pajahan]:
5 similar
[illegible]i sie se[ppo] Ilm[ollinen],
55 55 Takoja iän [ikuinen],



Eipä löyä V[äinämöi]stä.

[back to poem] [CSV] [map]

Cluster

SKVR I1 1. [52] **Eipä löyä V[äinämöi]stä.**
skvr01100010
Viena — Jyskyjärvi
1835 Lönnrot, Elias

1. Kertovat runot
└1. Epiikka
├ Kanteleen soitto
├ Kanteleen synty
├ Maailmansyntyruno
├ Sammon taonta
└ Väinämöisen ammunta

SKVR I1 2. [123] **Eibä löuvä Väinämöistä. -**
skvr01100020
Viena — Jyskyjärvi
1872 Borenius, A. A.

1. Kertovat runot
└1. Epiikka
├ Kanteleen soitto
├ Kanteleen synty
├ Kilpalaulanta
├ Sammon taonta
└ Väinämöisen ammunta

SKVR I1 16. [10] **Eipä löytty Väinämöistä.**
skvr01100160
Viena — Jyskyjärvi
1889 Meriläinen, Heikki

1. Kertovat runot
└1. Epiikka
├ Maailmansyntyruno
└ Väinämöisen ammunta
7. Erilaisia pieniä runoja
└1. Sananlaskuja
└ Jumalassa on jaoksun määrä, eikä miehen miehuessa

SKVR I1 58. [45] **Eipä löütüt Väinämöistä.**
skvr01100580
Viena — Vuokkiniemi
1871 Borenius, A. A.

1. Kertovat runot
└1. Epiikka
├ Maailmansyntyruno
├ Sammon ryöstö
├ Sammon taonta
└ Väinämöisen ammunta
2. Lyriikka
└02. Valituslaulut
└ 2. Laulut kodittomuudesta ja kulkemisesta
└ Vieraalle maalle joutunut

SKVR I1 58 a). [46] **45 Eipä löytyt V[äinä]möistä.**

[1] Maciej Janicki, Kati Kallio, and Mari Sarv. “Exploring Finnic written oral folk poetry through string similarity”. In: *Digital Scholarship in the Humanities* (in press).

METHODS

1. Verse similarity: bigram-based embedding + cosine similarity
2. **Poem similarity: weighted edit distance and alignment**

EDIT DISTANCE

Given two words, e.g. **plotting** and **poetry**, find the minimum number of *edit operations* needed to transform one word into the other.

An *edit operation* is a:

- insertion
- deletion
- substitution

of one character.

The optimal transformation is:

p	**l**	o	**t**	t	**i**	**n**	**g**
p		o	**e**	t	**r**	**y**	

with edit distance 5.

EDIT DISTANCE

		y_1 **p**	y_2 **o**	y_3 **e**	y_4 **t**	y_5 **r**	y_6 **y**
x_1	**p**						
x_2	**l**						
x_3	**o**						
x_4	**t**						
x_5	**t**						
x_6	**i**						
x_7	**n**						
x_8	**g**						

$$d_{i,j} = \min \begin{Bmatrix} d_{i-1,j} + 1 \\ d_{i,j-1} + 1 \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{Bmatrix}$$

where:

$$\delta(x_i, y_j) = \begin{cases} 0 \text{ if } x_i = y_j \\ 1 \text{ otherwise} \end{cases}$$

EDIT DISTANCE

			y_1	y_2	y_3	y_4	y_5	y_6
			p	**o**	**e**	**t**	**r**	**y**
		0	1	2	3	4	5	6
x_1	**p**	1						
x_2	**l**	2						
x_3	**o**	3						
x_4	**t**	4						
x_5	**t**	5						
x_6	**i**	6						
x_7	**n**	7						
x_8	**g**	8						

$$d_{i,j} = \min \begin{Bmatrix} d_{i-1,j} + 1 \\ d_{i,j-1} + 1 \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{Bmatrix}$$

where:

$$\delta(x_i, y_j) = \begin{cases} 0 \text{ if } x_i = y_j \\ 1 \text{ otherwise} \end{cases}$$

EDIT DISTANCE

			y_1	y_2	y_3	y_4	y_5	y_6
			p	**o**	**e**	**t**	**r**	**y**
		0	1	2	3	4	5	6
x_1	**p**	1	0					
x_2	**l**	2						
x_3	**o**	3						
x_4	**t**	4						
x_5	**t**	5						
x_6	**i**	6						
x_7	**n**	7						
x_8	**g**	8						

$$d_{i,j} = \min \begin{Bmatrix} d_{i-1,j} + 1 \\ d_{i,j-1} + 1 \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{Bmatrix}$$

where:

$$\delta(x_i, y_j) = \begin{cases} 0 \text{ if } x_i = y_j \\ 1 \text{ otherwise} \end{cases}$$

EDIT DISTANCE

			y_1	y_2	y_3	y_4	y_5	y_6
			p	**o**	**e**	**t**	**r**	**y**
		0	1	2	3	4	5	6
x_1	**p**	1	0	1	2	3	4	5
x_2	**l**	2						
x_3	**o**	3						
x_4	**t**	4						
x_5	**t**	5						
x_6	**i**	6						
x_7	**n**	7						
x_8	**g**	8						

$$d_{i,j} = \min \begin{Bmatrix} d_{i-1,j} + 1 \\ d_{i,j-1} + 1 \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{Bmatrix}$$

where:

$$\delta(x_i, y_j) = \begin{cases} 0 \text{ if } x_i = y_j \\ 1 \text{ otherwise} \end{cases}$$

EDIT DISTANCE

			y_1	y_2	y_3	y_4	y_5	y_6
			p	**o**	**e**	**t**	**r**	**y**
		0	1	2	3	4	5	6
x_1	**p**	1	0	1	2	3	4	5
x_2	**l**	2	1	1	2	3	4	5
x_3	**o**	3						
x_4	**t**	4						
x_5	**t**	5						
x_6	**i**	6						
x_7	**n**	7						
x_8	**g**	8						

$$d_{i,j} = \min \begin{Bmatrix} d_{i-1,j} + 1 \\ d_{i,j-1} + 1 \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{Bmatrix}$$

where:

$$\delta(x_i, y_j) = \begin{cases} 0 \text{ if } x_i = y_j \\ 1 \text{ otherwise} \end{cases}$$

EDIT DISTANCE

			y_1	y_2	y_3	y_4	y_5	y_6
			p	**o**	**e**	**t**	**r**	**y**
		0	1	2	3	4	5	6
x_1	**p**	1	0	1	2	3	4	5
x_2	**l**	2	1	1	2	3	4	5
x_3	**o**	3	2	1	2	3	4	5
x_4	**t**	4						
x_5	**t**	5						
x_6	**i**	6						
x_7	**n**	7						
x_8	**g**	8						

$$d_{i,j} = \min \begin{Bmatrix} d_{i-1,j} + 1 \\ d_{i,j-1} + 1 \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{Bmatrix}$$

where:

$$\delta(x_i, y_j) = \begin{cases} 0 \text{ if } x_i = y_j \\ 1 \text{ otherwise} \end{cases}$$

EDIT DISTANCE

			y_1	y_2	y_3	y_4	y_5	y_6
			p	**o**	**e**	**t**	**r**	**y**
		0	1	2	3	4	5	6
x_1	**p**	1	0	1	2	3	4	5
x_2	**l**	2	1	1	2	3	4	5
x_3	**o**	3	2	1	2	3	4	5
x_4	**t**	4	3	2	2	2	3	4
x_5	**t**	5						
x_6	**i**	6						
x_7	**n**	7						
x_8	**g**	8						

$$d_{i,j} = \min \begin{Bmatrix} d_{i-1,j} + 1 \\ d_{i,j-1} + 1 \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{Bmatrix}$$

where:

$$\delta(x_i, y_j) = \begin{cases} 0 \text{ if } x_i = y_j \\ 1 \text{ otherwise} \end{cases}$$

EDIT DISTANCE

			y_1	y_2	y_3	y_4	y_5	y_6
			p	**o**	**e**	**t**	**r**	**y**
		0	1	2	3	4	5	6
x_1	**p**	1	0	1	2	3	4	5
x_2	**l**	2	1	1	2	3	4	5
x_3	**o**	3	2	1	2	3	4	5
x_4	**t**	4	3	2	2	2	3	4
x_5	**t**	5	4	3	3	2	3	4
x_6	**i**	6						
x_7	**n**	7						
x_8	**g**	8						

$$d_{i,j} = \min \begin{Bmatrix} d_{i-1,j} + 1 \\ d_{i,j-1} + 1 \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{Bmatrix}$$

where:

$$\delta(x_i, y_j) = \begin{cases} 0 \text{ if } x_i = y_j \\ 1 \text{ otherwise} \end{cases}$$

EDIT DISTANCE

			y_1	y_2	y_3	y_4	y_5	y_6
			p	**o**	**e**	**t**	**r**	**y**
		0	1	2	3	4	5	6
x_1	**p**	1	0	1	2	3	4	5
x_2	**l**	2	1	1	2	3	4	5
x_3	**o**	3	2	1	2	3	4	5
x_4	**t**	4	3	2	2	2	3	4
x_5	**t**	5	4	3	3	2	3	4
x_6	**i**	6	5	4	4	3	3	4
x_7	**n**	7						
x_8	**g**	8						

$$d_{i,j} = \min \begin{Bmatrix} d_{i-1,j} + 1 \\ d_{i,j-1} + 1 \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{Bmatrix}$$

where:

$$\delta(x_i, y_j) = \begin{cases} 0 \text{ if } x_i = y_j \\ 1 \text{ otherwise} \end{cases}$$

EDIT DISTANCE

			y_1	y_2	y_3	y_4	y_5	y_6
			p	**o**	**e**	**t**	**r**	**y**
		0	1	2	3	4	5	6
x_1	**p**	1	0	1	2	3	4	5
x_2	**l**	2	1	1	2	3	4	5
x_3	**o**	3	2	1	2	3	4	5
x_4	**t**	4	3	2	2	2	3	4
x_5	**t**	5	4	3	3	2	3	4
x_6	**i**	6	5	4	4	3	3	4
x_7	**n**	7	6	5	5	4	4	4
x_8	**g**	8						

$$d_{i,j} = \min \begin{Bmatrix} d_{i-1,j} + 1 \\ d_{i,j-1} + 1 \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{Bmatrix}$$

where:

$$\delta(x_i, y_j) = \begin{cases} 0 \text{ if } x_i = y_j \\ 1 \text{ otherwise} \end{cases}$$

EDIT DISTANCE

			y_1	y_2	y_3	y_4	y_5	y_6
			p	**o**	**e**	**t**	**r**	**y**
		0	1	2	3	4	5	6
x_1	**p**	1	0	1	2	3	4	5
x_2	**l**	2	1	1	2	3	4	5
x_3	**o**	3	2	1	2	3	4	5
x_4	**t**	4	3	2	2	2	3	4
x_5	**t**	5	4	3	3	2	3	4
x_6	**i**	6	5	4	4	3	3	4
x_7	**n**	7	6	5	5	4	4	4
x_8	**g**	8	7	6	6	5	5	5

$$d_{i,j} = \min \begin{Bmatrix} d_{i-1,j} + 1 \\ d_{i,j-1} + 1 \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{Bmatrix}$$

where:

$$\delta(x_i, y_j) = \begin{cases} 0 \text{ if } x_i = y_j \\ 1 \text{ otherwise} \end{cases}$$

EDIT DISTANCE

			y_1 **p**	y_2 **o**	y_3 **e**	y_4 **t**	y_5 **r**	y_6 **y**
		0	←	←	←	←	←	←
x_1	**p**	↑	↖	←	←	←	←	←
x_2	**l**	↑	↑	↖	←	←	←	←
x_3	**o**	↑	↑	↖	↖	←	←	←
x_4	**t**	↑	↑	↑	↖	↖	←	←
x_5	**t**	↑	↑	↑	↑	↖	←	←
x_6	**i**	↑	↑	↑	↑	↑	↖	←
x_7	**n**	↑	↑	↑	↑	↑	↑	↖
x_8	**g**	↑	↑	↑	↑	↑	↑	↑

$$d_{i,j} = \min \left\{ \begin{matrix} d_{i-1,j} + 1 \\ d_{i,j-1} + 1 \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{matrix} \right\} \begin{matrix} \uparrow \\ \leftarrow \\ \nwarrow \end{matrix}$$

where:

$$\delta(x_i, y_j) = \begin{cases} 0 \text{ if } x_i = y_j \\ 1 \text{ otherwise} \end{cases}$$

p **l** o **t** t **i** **n** **g**
p o **e** t **r** **y**

EDIT DISTANCE

			y_1 **p**	y_2 **o**	y_3 **e**	y_4 **t**	y_5 **r**	y_6 **y**
		0	←	←	←	←	←	←
x_1	**p**	↑	↖	←	←	←	←	←
x_2	**l**	↑	↑	↖	←	←	←	←
x_3	**o**	↑	↑	↖	↖	←	←	←
x_4	**t**	↑	↑	↑	↖	↖	←	←
x_5	**t**	↑	↑	↑	↑	↖	←	←
x_6	**i**	↑	↑	↑	↑	↑	↖	←
x_7	**n**	↑	↑	↑	↑	↑	↑	↖
x_8	**g**	↑	↑	↑	↑	↑	↑	↑

$$d_{i,j} = \min \left\{ \begin{matrix} d_{i-1,j} + 1 \\ d_{i,j-1} + 1 \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{matrix} \right\} \begin{matrix} \uparrow \\ \leftarrow \\ \nwarrow \end{matrix}$$

where:

$$\delta(x_i, y_j) = \begin{cases} 0 \text{ if } x_i = y_j \\ 1 \text{ otherwise} \end{cases}$$

p	**l**	o	**t**	t	**i**	**n**	**g**	
p		o	**e**	t	**r**	**y**		

MAXIMUM-WEIGHT ALIGNMENT

Similar as edit distance, but:

- insertions and deletions have weight 0,
- $\delta(x_i, y_j)$ is the **similarity** between x_i and y_j (between 0 and 1),
- we are looking for an alignment with **maximum** total weight.

For example, let:

$$\delta(x_i, y_j) = \begin{cases} 1 & \text{if } x_i = y_j \\ 0.8 & \text{for } (i, y) \text{ and } (r, l) \\ 0.5 & \text{for } (t, r), (t, l) \text{ and } (e, i) \\ 0 & \text{otherwise} \end{cases}$$

The optimal transformation is:

p	l	o		t	t	i	n	g	
p		o	e	t	r	y			

with total weight 4.3

MAXIMUM-WEIGHT ALIGNMENT

The optimal transformation is:

p	l	o		t	t	i	n	g
p		o	e	t	r	y		

with total weight 4.3

$$
\begin{aligned}
s_{\text{raw}} &= 4.3 && \text{– raw sim.} \\
s_l &= \frac{s_{\text{raw}}}{|\mathbf{x}|} = \frac{4.3}{8} \approx 0.54 && \text{– left-normalized sim.} \\
s_r &= \frac{s_{\text{raw}}}{|\mathbf{y}|} = \frac{4.3}{6} \approx 0.72 && \text{– right-normalized sim.} \\
s &= \frac{2}{\frac{1}{s_l} + \frac{1}{s_r}} = \frac{2 \cdot s_{\text{raw}}}{|\mathbf{x}| + |\mathbf{y}|} \approx 0.61 && \text{– avg. normalized sim.}
\end{aligned}
$$

	yht_ei_kuttsun_lemmingäistä	rujot_ne_reillä_reissuaabi	rammat_rattsahin_ajeli	sogiat_venozin_soudi	lemmingäin_on_poiga_piilo	pillojah_on_piilemässä	pahojah_pagenemassa
kuttsu_verisogeat							
ruiot_reellä_rembutteli							
rammat_rattsahin_ajeli			1				
sogeat_venosin_souti				.71			
yht_ei_kuttsu_lemmingästä	.88						
lemmingäne_on_piilopoiga					.84	.52	
piiloja_on_piilemässä					.57	.86	
pahoja_pagenemassa							.93

$$d_{i,j} = \max \left\{ \begin{matrix} d_{i-1,j} \\ d_{i,j-1} \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{matrix} \right\}$$

		yht_ei_kuttsun_lemmingäistä	rujot_ne_reillä_reissuaabi	rammat_rattsahin_ajeli	sogiat_venozin_soudi	lemmingäin_on_poiga_piilo	pillojah_on_piilemässä	pahojah_pagenemassa
	0	0	0	0	0	0	0	0
kuttsu_verisogeat	0							
ruiot_reellä_rembutteli	0							
rammat_rattsahin_ajeli	0			1				
sogeat_venosin_souti	0				.71			
yht_ei_kuttsu_lemmingästä	0	.88						
lemmingäne_on_piilopoiga	0					.84	.52	
piiloja_on_piilemässä	0					.57	.86	
pahoja_pagenemassa	0							.93

$$d_{i,j} = \max \left\{ \begin{array}{c} d_{i-1,j} \\ d_{i,j-1} \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{array} \right\}$$

		yht_ei_kuttsun_lemmingäistä	rujot_ne_reillä_reissuaabi	rammat_rattsahin_ajeli	sogiat_venozin_soudi	lemmingäin_on_poiga_piilo	pillojah_on_piilemässä	pahojah_pagenemassa
	0	0	0	0	0	0	0	0
kuttsu_verisogeat	0	0	0	0	0	0	0	0
ruiot_reellä_rembutteli	0							
rammat_rattsahin_ajeli	0			1				
sogeat_venosin_souti	0				.71			
yht_ei_kuttsu_lemmingästä	0	.88						
lemmingäne_on_piilopoiga	0					.84	.52	
piiloja_on_piilemässä	0					.57	.86	
pahoja_pagenemassa	0							.93

$$d_{i,j} = \max \left\{ \begin{array}{c} d_{i-1,j} \\ d_{i,j-1} \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{array} \right\}$$

		yht_ei_kuttsun_lemmingäistä	rujot_ne_reillä_reissuaabi	rammat_rattsahin_ajeli	sogiat_venozin_soudi	lemmingäin_on_poiga_piilo	pillojah_on_piilemässä	pahojah_pagenemassa
	0	0	0	0	0	0	0	0
kuttsu_verisogeat	0	0	0	0	0	0	0	0
ruiot_reellä_rembutteli	0	0	0	0	0	0	0	0
rammat_rattsahin_ajeli	0			1				
sogeat_venosin_souti	0				.71			
yht_ei_kuttsu_lemmingästä	0	.88						
lemmingäne_on_piilopoiga	0					.84	.52	
piiloja_on_piilemässä	0					.57	.86	
pahoja_pagenemassa	0							.93

$$d_{i,j} = \max \left\{ \begin{array}{c} d_{i-1,j} \\ d_{i,j-1} \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{array} \right\}$$

		yht_ei_kuttsun_lemmingäistä	rujot_ne_reillä_reissuaabi	rammat_rattsahin_ajeli	sogiat_venozin_soudi	lemmingäin_on_poiga_piilo	pillojah_on_piilemässä	pahojah_pagenemassa
	0	0	0	0	0	0	0	0
kuttsu_verisogeat	0	0	0	0	0	0	0	0
ruiot_reellä_rembutteli	0	0	0	0	0	0	0	0
rammat_rattsahin_ajeli	0	0	0	1	1	1	1	1
sogeat_venosin_souti	0				.71			
yht_ei_kuttsu_lemmingästä	0	.88						
lemmingäne_on_piilopoiga	0						.52	
piiloja_on_piilemässä	0					.57	.86	
pahoja_pagenemassa	0							.93

$$d_{i,j} = \max \left\{ \begin{array}{c} d_{i-1,j} \\ d_{i,j-1} \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{array} \right\}$$

		yht_ei_kuttsun_lemmingäistä	rujot_ne_reillä_reissuaabi	rammat_rattsahin_ajeli	sogiat_venozin_soudi	lemmingäin_on_poiga_piilo	pillojah_on_piilemässä	pahojah_pagenemassa
	0	0	0	0	0	0	0	0
kuttsu_verisogeat	0	0	0	0	0	0	0	0
ruiot_reellä_rembutteli	0	0	0	0	0	0	0	0
rammat_rattsahin_ajeli	0	0	0	1	1	1	1	1
sogeat_venosin_souti	0	0	0	1	1.71	1.71	1.71	1.71
yht_ei_kuttsu_lemmingästä	0	.88						
lemmingäne_on_piilopoiga	0						.52	
piiloja_on_piilemässä	0						.86	
pahoja_pagenemassa	0							.93

$$d_{i,j} = \max \left\{ \begin{array}{c} d_{i-1,j} \\ d_{i,j-1} \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{array} \right\}$$

		yht_ei_kuttsun_lemmingäistä	rujot_ne_reillä_reissuaabi	rammat_rattsahin_ajeli	sogiat_venozin_soudi	lemmingäin_on_poiga_piilo	pillojah_on_piilemässä	pahojah_pagenemassa
	0	0	0	0	0	0	0	0
kuttsu_verisogeat	0	0	0	0	0	0	0	0
ruiot_reellä_rembutteli	0	0	0	0	0	0	0	0
rammat_rattsahin_ajeli	0	0	0	1	1	1	1	1
sogeat_venosin_souti	0	0	0	1	1.71	1.71	1.71	1.71
yht_ei_kuttsu_lemmingästä	0	.88	.88	1	1.71	1.71	1.71	1.71
lemmingäne_on_piilopoiga	0						.52	
piiloja_on_piilemässä	0						.86	
pahoja_pagenemassa	0							.93

$$d_{i,j} = \max \left\{ \begin{array}{c} d_{i-1,j} \\ d_{i,j-1} \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{array} \right\}$$

		yht_ei_kuttsun_lemmingäistä	rujot_ne_reillä_reissuaabi	rammat_rattsahin_ajeli	sogiat_venozin_soudi	lemmingäin_on_poiga_piilo	piilojah_on_piilemässä	pahojah_pagenemassa
	0	0	0	0	0	0	0	0
kuttsu_verisogeat	0	0	0	0	0	0	0	0
ruiot_reellä_rembutteli	0	0	0	0	0	0	0	0
rammat_rattsahin_ajeli	0	0	0	1	1	1	1	1
sogeat_venosin_souti	0	0	0	1	1.71	1.71	1.71	1.71
yht_ei_kuttsu_lemmingästä	0	.88	.88	1	1.71	1.71	1.71	1.71
lemmingäne_on_piilopoiga	0	.88	.88	1	1.71	2.55	2.55	2.55
piiloja_on_piilemässä	0						.86	
pahoja_pagenemassa	0							.93

$$d_{i,j} = \max \left\{ \begin{array}{c} d_{i-1,j} \\ d_{i,j-1} \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{array} \right\}$$

		yht_ei_kuttsun_lemmingäistä	rujot_ne_reillä_reissuaabi	rammat_rattsahin_ajeli	sogiat_venozin_soudi	lemmingäin_on_poiga_piilo	pillojah_on_piilemässä	pahojah_pagenemassa
	0	0	0	0	0	0	0	0
kuttsu_verisogeat	0	0	0	0	0	0	0	0
ruiot_reellä_rembutteli	0	0	0	0	0	0	0	0
rammat_rattsahin_ajeli	0	0	0	1	1	1	1	1
sogeat_venosin_souti	0	0	0	1	1.71	1.71	1.71	1.71
yht_ei_kuttsu_lemmingästä	0	.88	.88	1	1.71	1.71	1.71	1.71
lemmingäne_on_piilopoiga	0	.88	.88	1	1.71	2.55	2.55	2.55
piiloja_on_piilemässä	0	.88	.88	1	1.71	2.55	3.41	3.41
pahoja_pagenemassa	0							.93

$$d_{i,j} = \max \left\{ \begin{array}{c} d_{i-1,j} \\ d_{i,j-1} \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{array} \right\}$$

		yht_ei_kuttsun_lemmingäistä	rujot_ne_reillä_reissuaabi	rammat_rattsahin_ajeli	sogiat_venozin_soudi	lemmingäin_on_poiga_piilo	pillojah_on_piilemässä	pahojah_pagenemassa
	0	0	0	0	0	0	0	0
kuttsu_verisogeat	0	0	0	0	0	0	0	0
ruiot_reellä_rembutteli	0	0	0	0	0	0	0	0
rammat_rattsahin_ajeli	0	0	0	1	1	1	1	1
sogeat_venosin_souti	0	0	0	1	1.71	1.71	1.71	1.71
yht_ei_kuttsu_lemmingästä	0	.88	.88	1	1.71	1.71	1.71	1.71
lemmingäne_on_piilopoiga	0	.88	.88	1	1.71	2.55	2.55	2.55
piiloja_on_piilemässä	0	.88	.88	1	1.71	2.55	3.41	3.41
pahoja_pagenemassa	0	.88	.88	1	1.71	2.55	3.41	4.34

$$d_{i,j} = \max \left\{ \begin{array}{c} d_{i-1,j} \\ d_{i,j-1} \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{array} \right\}$$

		yht_ei_kuttsun_lemmingäistä	rujot_ne_reillä_reissuaabi	rammat_rattsahin_ajeli	sogiat_venozin_soudi	lemmingäin_on_poiga_piilo	pillojah_on_piilemässä	pahojah_pagenemassa
	0	0	0	0	0	0	0	0
kuttsu_verisogeat	0	0	0	0	0	0	0	0
ruiot_reellä_rembutteli	0	0	0	0	0	0	0	0
rammat_rattsahin_ajeli	0	0	0	1	1	1	1	1
sogeat_venosin_souti	0	0	0	1	1.71	1.71	1.71	1.71
yht_ei_kuttsu_lemmingästä	0	.88	.88	1	1.71	1.71	1.71	1.71
lemmingäne_on_piilopoiga	0	.88	.88	1	1.71	2.55	2.55	2.55
piiloja_on_piilemässä	0	.88	.88	1	1.71	2.55	3.41	3.41
pahoja_pagenemassa	0	.88	.88	1	1.71	2.55	3.41	4.34

$$d_{i,j} = \max \begin{Bmatrix} d_{i-1,j} \\ d_{i,j-1} \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{Bmatrix}$$

$$s_{\text{raw}} = 4.34$$

$$s_l = \frac{4.34}{8} \approx 0.54$$

$$s_r = \frac{4.34}{7} = 0.62$$

$$s = \frac{2 \cdot 4.34}{8 + 7} \approx 0.58$$

$$d_{i,j} = \max \left\{ \begin{matrix} d_{i-1,j} \\ d_{i,j-1} \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{matrix} \right\} \begin{matrix} \uparrow \\ \leftarrow \\ \nwarrow \end{matrix}$$

	0	yht_ei_kuttsun_lemmingäistä	rujot_ne_reillä_reissuaabi	rammat_rattsahin_ajeli	sogiat_venozin_soudi	lemmingäin_on_poiga_piilo	pillojah_on_piilemässä	pahojah_pagenemassa
	0	←	←	←	←	←	←	←
kuttsu_verisogeat	↑	↑	↑	↑	↑	↑	↑	↑
ruiot_reellä_rembutteli	↑	↑	↑	↑	↑	↑	↑	↑
rammat_rattsahin_ajeli	↑	↑	↑	↖	←	←	←	←
sogeat_venosin_souti	↑	↑	↑	↑	↖	←	←	←
yht_ei_kuttsu_lemmingästä	↑	↖	←	↑	↑	↑	↑	↑
lemmingäne_on_piilopoiga	↑	↑	↑	↑	↑	↖	←	←
piiloja_on_piilemässä	↑	↑	↑	↑	↑	↑	↖	←
pahoja_pagenemassa	↑	↑	↑	↑	↑	↑	↑	↖

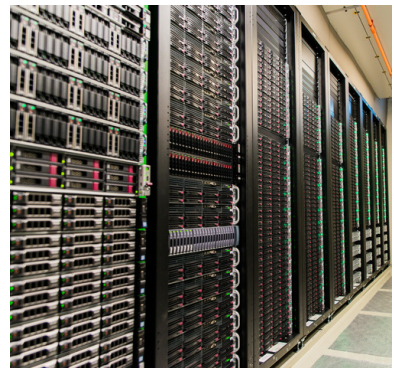
$$\begin{aligned} s_{\text{raw}} &= 4.34 \\ s_l &= \frac{4.34}{8} \approx 0.54 \\ s_r &= \frac{4.34}{7} = 0.62 \\ s &= \frac{2 \cdot 4.34}{8+7} \approx 0.58 \end{aligned}$$

$$d_{i,j} = \max \left\{ \begin{matrix} d_{i-1,j} \\ d_{i,j-1} \\ d_{i-1,j-1} + \delta(x_i, y_j) \end{matrix} \right\} \begin{matrix} \uparrow \\ \leftarrow \\ \nwarrow \end{matrix}$$

$$\begin{aligned} s_{\text{raw}} &= 4.34 \\ s_l &= \frac{4.34}{8} \approx 0.54 \\ s_r &= \frac{4.34}{7} = 0.62 \\ s &= \frac{2 \cdot 4.34}{8+7} \approx 0.58 \end{aligned}$$

Maximum-weight alignment

- compute on all pairs from among 275,181 poems
- optimization: use vector and matrix operations
  - compute entire rows at once
  - compute the alignment between one poem and all others at once
- run on a GPU
  (a $13,000 NVidia Tesla V100 with 32 GB memory)
- done within 64h!
- result: over 16 million pairs of similar poems
- criteria: $s_{\text{raw}} > 2$ and $\min\{s_l, s_r\} > 0.1$



RESULT: RUNOREGI

SKVR VII1 803.

skvr07108030
Laatokan Karjala (Raja-Karjala) — Suistamo
1897 Borenius, A. A.

Metadata

COL Bor.
ID 803.
INF Muuanto. Iivana Shemeikka ("Jehkin Iivana"), 57 v.
LOC Suistamo.
OSA VII1
SGN III A, s. 304.
TMP -7/6 97.

Themes

1. Kertovat runot
└ 1. Epikka
└ Lemminkäisen virsi

Similar poems

[compare all]

SKVR VII1 808.	34 %	1884 Laatokan Karjala (Raja-Karjala) — Suistamo Basilier, Hj
SKVR VII1 806.	34 %	1884 Laatokan Karjala (Raja-Karjala) — Suistamo Basilier, Hj
SKVR VII1 806 b.	32 %	1884 Laatokan Karjala (Raja-Karjala) — Suistamo Krohn, Kaarle
SKVR VII1 804.	29 %	1897 Laatokan Karjala (Raja-Karjala) — Suistamo Borenius, A. A.
SKVR VII1 805.	27 %	1900 Laatokan Karjala (Raja-Karjala) — Suistamo Härkönen, Iivo
SKVR VII1 809.	27 %	1915 Laatokan Karjala (Raja-Karjala) — Suistamo Hannikainen, Lauri
SKVR VII1 814.	27 %	1897 Laatokan Karjala (Raja-Karjala) — Korpiselkä Potschtareff, Wasilei
SKVR VII1 803 a.	26 %	1888 Laatokan Karjala (Raja-Karjala) — Suistamo Relander, O.
JR 31254	26 %	1935 Laatokan Karjala (Raja-Karjala) — Suistamo Kähmi, Martta
SKVR VII1 816.	22 %	1900 Laatokan Karjala (Raja-Karjala) — Korpiselkä Härkönen, Iivo
SKVR VII1 779.	21 %	1884 Laatokan Karjala (Raja-Karjala) — Suistamo Krohn, Kaarle
SKVR VII1 793.	20 %	1884 Laatokan Karjala (Raja-Karjala) — Suistamo Basilier, Hj
JR 05073	20 %	1935 Laatokan Karjala (Raja-Karjala) — Suistamo Haavio, Martti
SKVR VII1 794.	19 %	1884 Laatokan Karjala (Raja-Karjala) — Suistamo Krohn, Kaarle
SKVR VII1 778.	19 %	1900 Laatokan Karjala (Raja-Karjala) — Suistamo Härkönen, Iivo
SKVR VII1 789.	19 %	1884 Laatokan Karjala (Raja-Karjala) — Suistamo Basilier, Hj
SKVR VII1 775.	19 %	1909 Laatokan Karjala (Raja-Karjala) — Suistamo Niemi, A. R.

Text

[show verse-level themes] [show shared verses matrix]

Lemmingäzes lauletah.

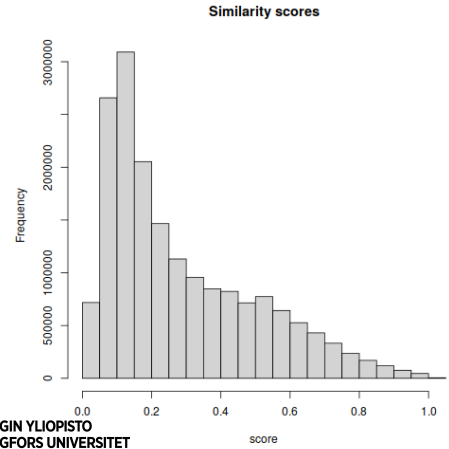
[2] Savupa soarelle palaabi,
[3] Niemembä kylgyvöt kydöibi.
[4] Toivoimba#1 paimozen tuleksi#2,#3,
[5] Suur' oli paimozen tuleksi#3;
[6] 5 Toivoih on#4 sodisavuksi,
[7] Pieñ oli#5 sodisavuksi.
[8] Osmatta on#6 olutta keitti,
[9] *Kallervoñiba#7 kul'l'oivetta*
[10] Yheksäss#8 ozranjyvässä,
[11] 10 Kaheksas#9 kagranjyvässä,
[12] Tuljiillaba vierahilla.
[13] *Laittobi vieštit viizjillä,
[14] Kutsutpa kuuzilla jagell*,
[15] Kuttšuba rujot, kuttsu rammat,
[16] 15 Kuttšubon perisogiat,
[17] Kuttšuba - - -
[18] Yht' ei kuttsun Lemmingäistä.
[19] *Rujot (ne) reillä reissuaabi,
[20] Rammat rattšahin ajeli,
[21] 20 Sogiat venozin soudi.*
[22] Lemmingäin on poiga piilo
[23] Pillojah on ptilemässä,#10
[24] Pahojah pagenemassa.#11
[25] "Hoib om moamo, kandajañi,
[26] 25 Armas maijon andajañi,
[27] Ihalan imettäjäni,
[28] Ettsib om miul pelvoi paid[a],
[29] Ennembä neidona kuvottu,
[30] Kassabapeän#12 on kalkuteltu,
[31] 30 Kannabas paloni paid[a]."
[32] "Minne (sie) lähtet, poiguvoñi?"
[33] "Lähtembä Väinölän pidoihe,
[34] Jumaliston juomingihe,
[35] Suuremba synnin syömingihe."
[36] 35 "(N)Eläs (sie) lähte, poiguvoñi#13,
[37] Sielä siul on 3 surmoa."

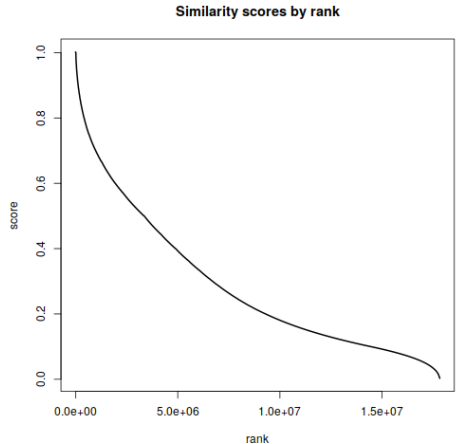
QUANTITATIVE ANALYSIS

Research questions:

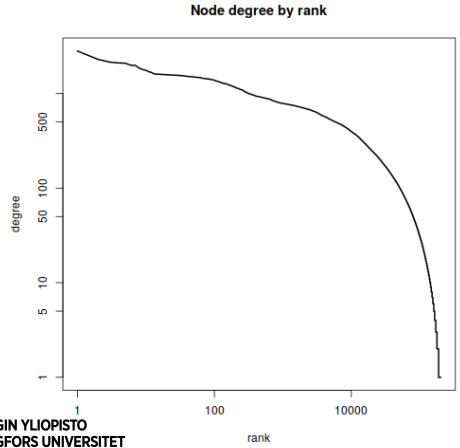
1. Types of relationships between similar texts
   same singer / parish / poem type? temporal and geographical distance?
2. Oral tradition $\leftrightarrow$ printed works
3. Building type indices
   does the similarity computation help improve and evaluate them?

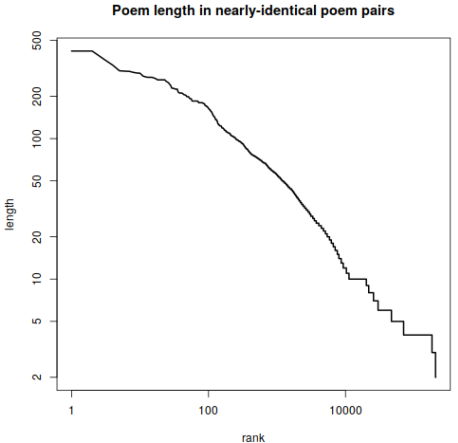
OVERVIEW


Similarity scores
Frequency
0
500000
1000000
2000000
3000000
0.0
0.2
0.4
0.6
0.8
1.0
score

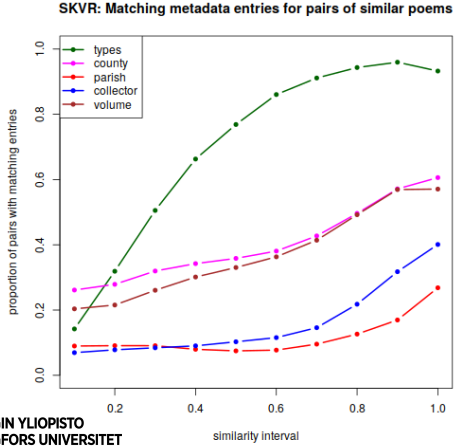

Similarity scores by rank
score
0.0
0.2
0.4
0.6
0.8
1.0
0.0e+00
5.0e+06
1.0e+07
1.5e+07
rank

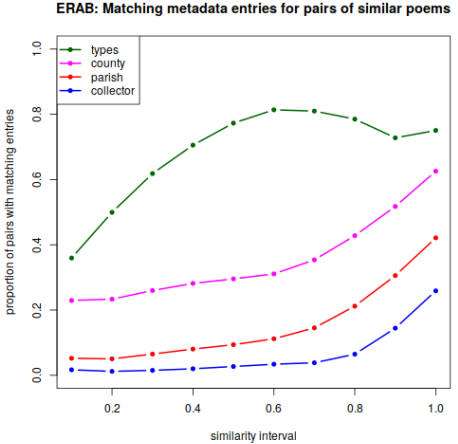
OVERVIEW


Node degree by rank
degree
1
5
10
50
100
500
1
100
10000
rank

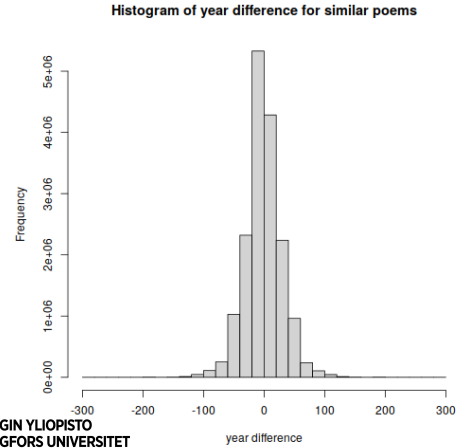

Poem length in nearly-identical poem pairs
length
2
5
10
20
50
100
200
500
1
100
10000
rank

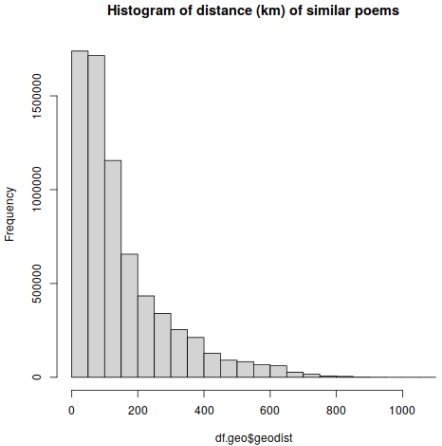
OVERVIEW


SKVR: Matching metadata entries for pairs of similar poems
types
county
parish
collector
volume
proportion of pairs with matching entries
similarity interval
0.0
0.2
0.4
0.6
0.8
1.0


ERAB: Matching metadata entries for pairs of similar poems
types
county
parish
collector
proportion of pairs with matching entries
similarity interval
0.0
0.2
0.4
0.6
0.8
1.0

OVERVIEW


Histogram of year difference for similar poems
Frequency
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
-300
-200
-100
0
100
200
300
year difference


Histogram of distance (km) of similar poems
Frequency
0
500000
1000000
1500000
0
200
400
600
800
1000
df.geo$geodist

OUTLIER ANALYSIS

Similar poems with the largest temporal distance:

- charms (e.g. 1 2 3 4 5)
- proverbs (e.g. 1 2 3)
- epics: ‘The spider song’ (1)

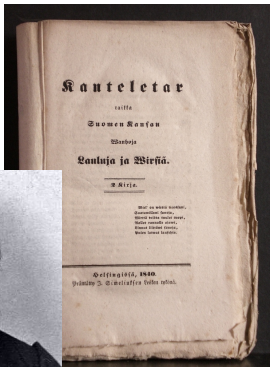
Similar poems with the largest geographic distance:

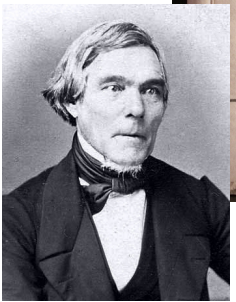
- charms (e.g. 1 2)
- lullabies (e.g. 1 2)
- children’s songs (e.g. 1 2)

ORAL-LITERARY RELATIONSHIPS

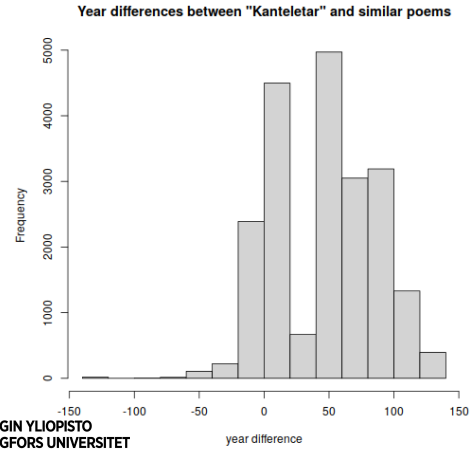
Example: Kanteletar

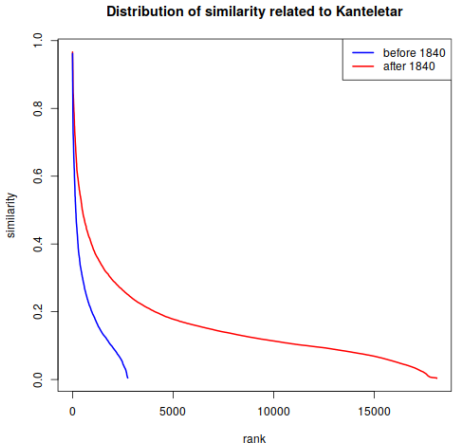
- a large collection of folk poems compiled by Elias Lönnrot
- 676 poems
- published in 1840
- close to oral tradition
- very well-known and influential



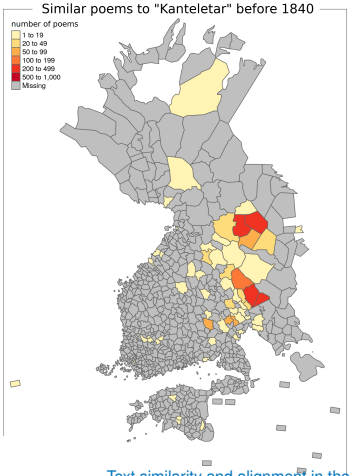


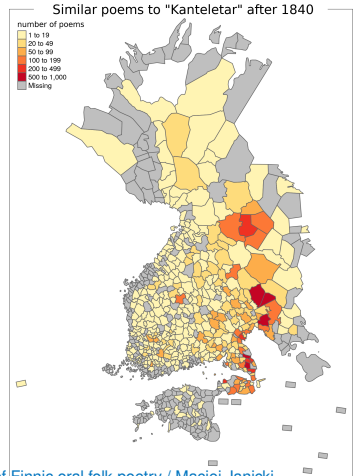
ORAL-LITERARY RELATIONSHIPS





ORAL-LITERARY RELATIONSHIPS





ORAL-LITERARY RELATIONSHIPS

Poems similar to “Kanteletar” by collector:

before 1840		
collector	**n**	**examples**
Lönnrot, Elias	1798	1 2 3 4 5
Gottlund, K. A.	207	1 2 3 4 5
Cajan, J. Fr.	185	1 2 3 4 5
Keckman, K. N.	128	1 2 3 4 5
Ganander, Christfrid	89	1 2 3 4 5
...		

1840 and after		
collector	**n**	**examples**
Europaeus, D. E. D.	1008	1 2 3 4 5
Alava, Vihtori	207	1 2 3 4 5
Ahlqvist A. E.	185	1 2 3 4 5
Lönnrot, Elias	128	1 2 3 4 5
Krohn, Kaarle	89	1 2 3 4 5
...		

ORAL-LITERARY RELATIONSHIPS

Poems similar to “Kanteletar” by singer:

before 1840		
collector	**n**	**examples**
Arhippa Perttuńe	78	1 2 3 4 5
Juhana Kainulainen	34	1 2 3 4 5
Martiska Karjalaińi	13	1 2 3 4 5
Matleena Kuivalatar	11	1 2 3 4 5
Mateli Kuivalatar	9	1 2 3 4 5
...		

1840 and after		
collector	**n**	**examples**
Larin Paraske	401	1 2 3 4 5
Liisa Itko	76	1 2 3 4 5
Arhippaińi Miihkali	46	1 2 3 4 5
Vyötär Tapanainen	45	1 2 3 4 5
Mari Laakko	35	1 2 3 4 5
...		

ENHANCING THE TYPE INDICES

Proposing type annotations for unannotated poems:

SKVR		
type	**n**	**examples**
The maid to be ransomed	19	1 2 3 4 5
Words of fire	16	1 2 3 4 5
Jacob de la Gardie	15	1 2 3 4 5
Words of dry weather	12	1 2 3 4 5
Golden maiden	12	1 2 3 4 5
...		
total	711	

ERAB		
type	**n**	**examples**
The Martinmas song	741	1 2 3 4 5
Marry me!	630	1 2 3 4 5
Harvesting song	369	1 2 3 4 5
Wirble-warble!	300	1 2 3 4 5
Song of communal work	314	1 2 3 4 5
...		
total	32806	

ENHANCING THE TYPE INDICES

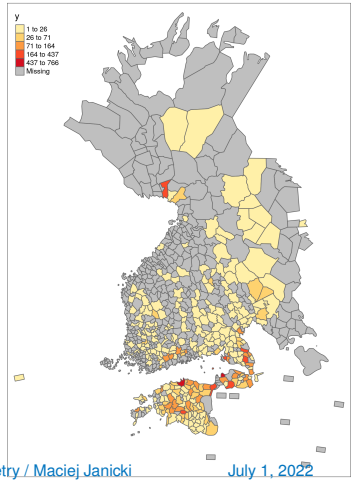
(Nearly) identical poems with different type annotations?

Type 1	Type 2	n	examples
Shoo, pig, to hunt foxes	I go to the forest to hunt foxes	70	1 2 3 4 5
The Lord will wake you up	Sleep, sleep my grassbird	57	1 2 3 4 5
Growing of hair	Wolves and sheep	36	1 2 3 4 5
Sing as long as you live	Sing as long as you live!	34	1 2 3 4 5
The sad one's home	Maiden's town	33	1 2 3 4 5
	...		

SIMILARITIES ACROSS COLLECTIONS

(table excludes ERAB collections from Ingria)

SKVR type	ERAB type	n	examples
Shoo, goat, to the cattle	Shoo, goat, to the cattle	780	1 2 3 4 5
The maid to be ransomed	The maid to be ransomed	412	1 2 3 4 5
I served the farmer my years	The taskmaster's farmhand	94	1 2 3 4 5
The thief takes the jewelry	A horse stolen	29	1 2 3 4 5
Singer's reward	Singer's reward	42	1 2 3 4 5
	...		



CONCLUSIONS

- *Relational* view on the collections
- Large-scale patterns have non-trivial justifications
  - different stability for different kinds of texts
  - collection density varies depending on time and place
  - influence of particular collectors / singers / works
- Desirable: a methodology **between** close and distant reading
- Proposed workflow:
  1. select poem pairs according to some criteria based on metadata,
  2. browse through aligned views of representative examples.
- Generalize? (characteristics: size of the collections, amount of similarity)