Introducing Relate

A Software for Establishing Quantitative Relationships between Manuscripts

Pasi Hyytiäinen
University of Helsinki

What is Relate?

A software that efficiently relates manuscripts in a quantitative manner

- **Uses transcriptions only**
  - Does not require collations, variation units or encodings
- **Once texts are transcribed, it can relate hundreds of manuscripts in matter of hours**
  - No need to rely on samples anymore
  - All textual data can be considered in each manuscript tradition
- **Now exists as source code**
  - Does not have GUI (graphical user interface)
- **Written in Python**
- **GitHub**

Background

Colwell and Tune

- **”Quantitative method of textual analysis”**
- **Based on collations and units of variation**
- **Records instances of agreements in places of variation**
  - Agreements are converted into percentages by dividing the number of agreements and the number of all variation places between pairs of MSS
- **Definition of a variation place**
  - A segment of a text containing at least two variants supported by at least two MSS
- **Genetically significant variants**

Gordon Fee

- **Calculating agreements in two stages**
- **Continued the usage of genetically significant variants**

Background

CBGM

- **Pre-genealogical coherence**
- **Singular readings are also considered**
- **Almost all types variations are considered**

The concept of a genetically significant variant lives on

- **Hurtado (1981), Geer (1994), Osburn (2004), Donker (2011)**

Problems

- **Different definitions of the variatio unit leads to differing agreements rates**
- **Critics divides texts into variation units very differently**
  - This also affects the agreement rates
  - Colwell and Tune acknowledged this problem already in 1964
- **Coventional quantitative methods of textual analysis takes lots of time**

A	the fox jumped over the hedge
B	-
C	the cat jumped over the fence
D	a man saw that the fox jumped over the hedge
E	a man saw that the fox jumped over the fence

Solution 1			Solution 2		
A	–	the fox jumped over the hedge	–	the fox jumped	over the hedge
B	–	–	–	–	–
C	–	the cat jumped over the fence	–	the cat jumped	over the fence
D	a man saw that	the fox jumped over the hedge	a man saw that	the fox jumped	over the hedge
E	a man saw that	the fox jumped over the fence	a man saw that	the fox jumped	over the fence

Agreement rates	
A, D	50 %
C, E	0 %

Agreement rates	
A, D	66,6 %
C, E	33,3 %

A	the fox jumped over the hedge
B	-
C	the cat jumped over the fence
D	a man saw that the fox jumped over the hedge
E	a man saw that the fox jumped over the fence

Solution 1			Solution 2		
A	–	the fox jumped over the hedge	–	the fox jumped	over the hedge
B	–			-	–
C	–			jumped	over the fence
D	a man saw tha			jumped	over the hedge
E	a man saw tha			jumped	over the fence

Countless of decisions needs to be made and every one of them have a direct impact on the similarity values

Agreement rates	
A, D	50 %
C, E	0 %

Agreement rates	
A, D	66,6 %
C, E	33,3 %

Toward a New Way of Thinking

Testing different stemmatological approaches

- **CBGM**
  - *”Evolving Gamaliel Tradition in Codex Bezae Cantabrigiensis, Acts 5:38–39: A Novel Application of Coherence-Based Genealogical Method (CBGM)”*
- **Phylomemetics**
  - *”The Changing Text of Acts: A Phylogenetic Approach”*

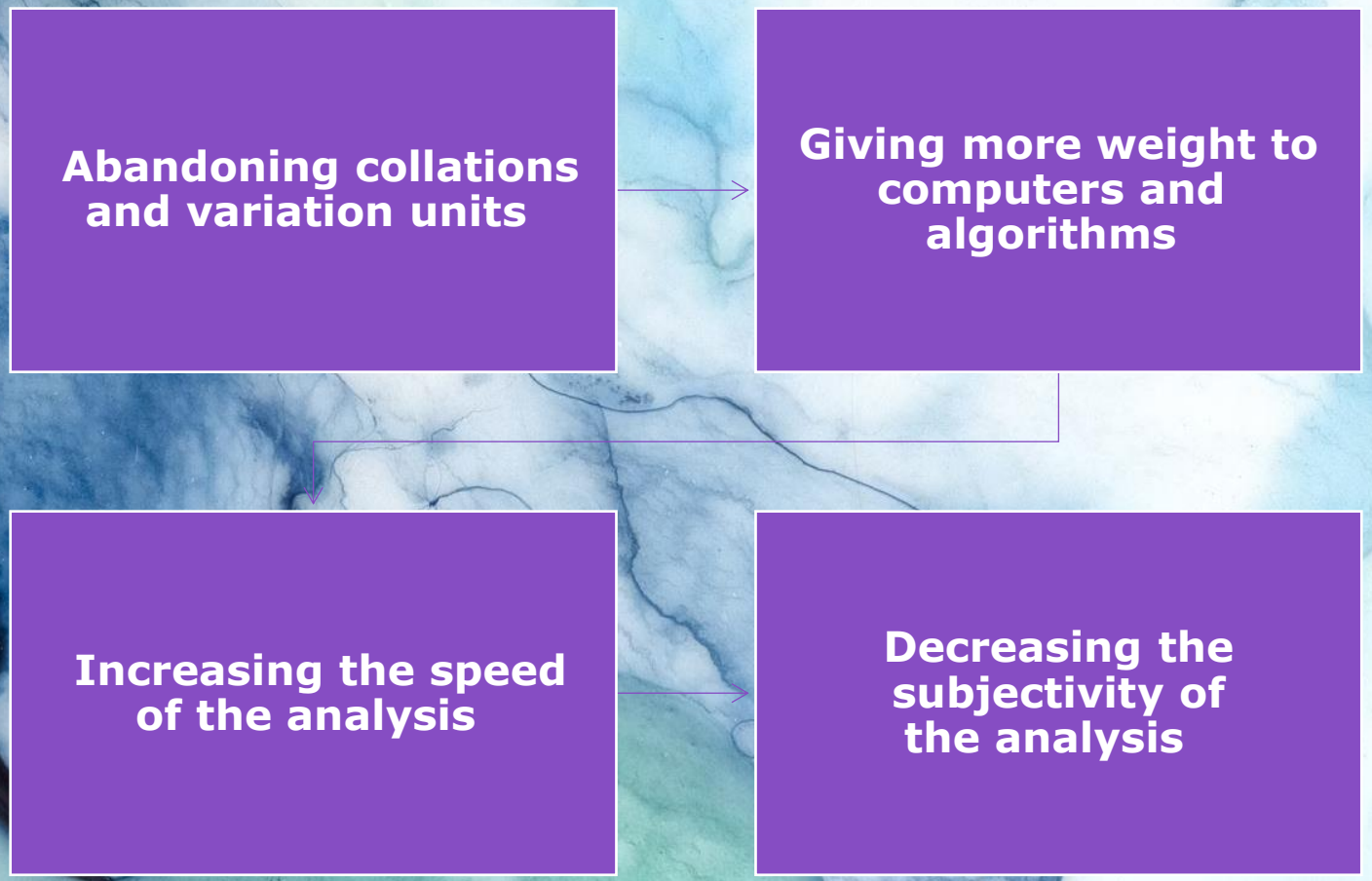
Interdisciplinary state of mind

- **String metrics**
- **Set theory**
- **Data mining**

Considering all textual data

- **Relying on samples is not an ideal situation**
  - *Teststellen*
  - Michelle Barbi on Dante
    - From 396 lines, only 121 proved to be useful

Toward a New Way of Thinking



Used Methodology

Theoretical foundation of the methodology:

- *”A new method in establishing quantitative relationships between manuscripts of the New Testament”* in *Digital scholarship in the Humanities, 2022* (open-access)

String metrics

- **Measures the similarity between two strings**
  - String is a sequence of chracters = a text
- **Allows one to operate using transcriptions only**
  - No need for collations or variation units
- **This can be calculated in several ways**
  - Character-based string metrics
    - *Levehnstein, Hamming, Jaro etc.*
  - Token-based string metrics
    - *Bag-of-Words (BOW)*
    - *Shingling*
    - *Jaccard, Overlap, and Sørensen-Dice*

Used Methodology

Problems with the character-based approaches

- **Cannot be used in Greek NT manuscripts**
  - Records similarities between words that are entirely different
    - περιβάλλω / περιπλέπω
  - Results to overly high similarity values
- **Requires too many operations**
  - Calculations takes too much time

Token-based approaches

- **Are much simpler**
  - Demand fewer operations
  - Faster
- **Record all types of variations between manuscripts**
  - Word changes, additions, deletions, word order changes
- **Results to similar agreement rates compared to conventional calculations**
- **Standardization of the spelling is recommended**

Step 1: K-Shingling

a man saw that the fox jumped over the fence

a man	man saw	saw that	that the	the fox	fox jumped	jumped over	over the	the fence

that the	fox jumped	the fence	a man	jumped over	over the	the fox	saw that	man saw

Step 1: K-Shingling

a man saw that the fox jumped over the fence

a man saw that the fox jumped over the hedge

fox jumped	man saw	jumped over	that the	saw that	the fox	a man	over the	the fence	
that the	fox jumped	the hedge	a man	jumped over	over the	the fox	saw that	man saw	

Step 2: Calculating similarities

Set1	fox jumped	man saw	jumped over	that the	saw that	the fox	a man	over the	the fence

Set2	that the	fox jumped	the hedge	a man	jumped over	over the	the fox	saw that	man saw

Word bigram	*Set 1*	*Set 2*
fox jumped	1	1
man saw	1	1
jumped over	1	1
that the	1	1
saw that	1	1
the fox	1	1
a man	1	1
over the	1	1
the hedge	0	1
the fence	1	0

Intersection = 8

Union = 10

Jaccard coefficient = $\frac{intersection}{union}$

= 8 / 10 = 0.8 = 80 %

Step 2: Calculating similarities

Set1	fox jumped	man saw	jumped over	that the	saw that	the fox	a man	over the	the fence

Set 2	that the	fox jumped	the hedge	a man	jumped over	over the	the fox	saw that	man saw

Word bigram	*Set 1*	*Set 2*
fox jumped	1	1
man saw	1	1
jumped over	1	1
that the	1	1
saw that	1	1
the fox	1	1
a man	1	1
over the	1	1
the hedge	0	1
the fence	1	0

Intersection = 8

Union = 10

Jaccard coefficient = $\frac{intersection}{union}$

= 8 / 10 = 0.8 = 80 %

Step 2: Calculating similarities

Set1	fox jumped	man saw	jumped over	that the	saw that	the fox	a man	over the	the fence

Set2	that the	fox jumped	the hedge	a man	jumped over	over the	the fox	saw that	man saw

Word bigram	*Set1*	*Set2*
fox jumped	1	1
man saw	1	1
jumped over	1	1
that the	1	1
saw that	1	1
the fox	1	1
a man	1	1
over the	1	1
the hedge	0	1
the fence	1	0

Intersection = 8

Union = 10

Sørensen-Dice Coefficient (SDC)

$$\frac{2\ x\ intersection}{sum\ of\ the\ number\ of\ elements\ in\ each\ set}$$

= 16 / 18 = 0.888 = 88 %

Letter bigram	*Set1*	*Set2*
'a	1	1
' m'	1	1
'ma'	1	1
'an'	1	1
'an'	1	1
'n '	1	1
' s'	1	1
'sa'	1	1
'aw'	1	1
'w '	1	1
' t'	1	1
'th'	1	1
'ha'	1	1
'at'	1	1
't '	1	1
' t'	1	1
'th'	1	1
'he'	1	1
'e '	1	1
' f'	1	1
'fo'	1	1
'ox'	1	1

Sørensen-Dice Coefficient
64 / 71 = 0.90 = 90 %

Speed

Token-based approach is fast

- **The combination of k-shingling and the Sørensen-Dice coefficient**
  - 54 manuscripts of Acts can be analyzed, using the letter-grams, in their entirety (28 chapters) in ten minutes
    - (2916 comparisons x 0.21 sec = 612 sec = 10.2 min)

Character-based approach is slower

- **Fastest Levenshtein algorithm (Myers)**
  - 54 manuscripts of Acts can be analyzed in their entirety in 100 minutes
    - (2916 comparisons x 2,05 sec = 6000 sec = 100 min)

Accuracy: Acts 5

MSS	CBGM	K-shingling + SDC
03, 05	73.93	74.12
03, 1175	94.26	94.26
614, 876	89.54	90.84
1409, 1739	91.24	91.71

Relate includes

Tokenization:

- K-shingling (character and word)

Character-based metrics:

- Levenshtein and Hamming

Token-based metrics:

- Jaccard, Overlap, Sørensen-Dice

Matrixes:

- Similarity, distance (with or without a standardizing function)

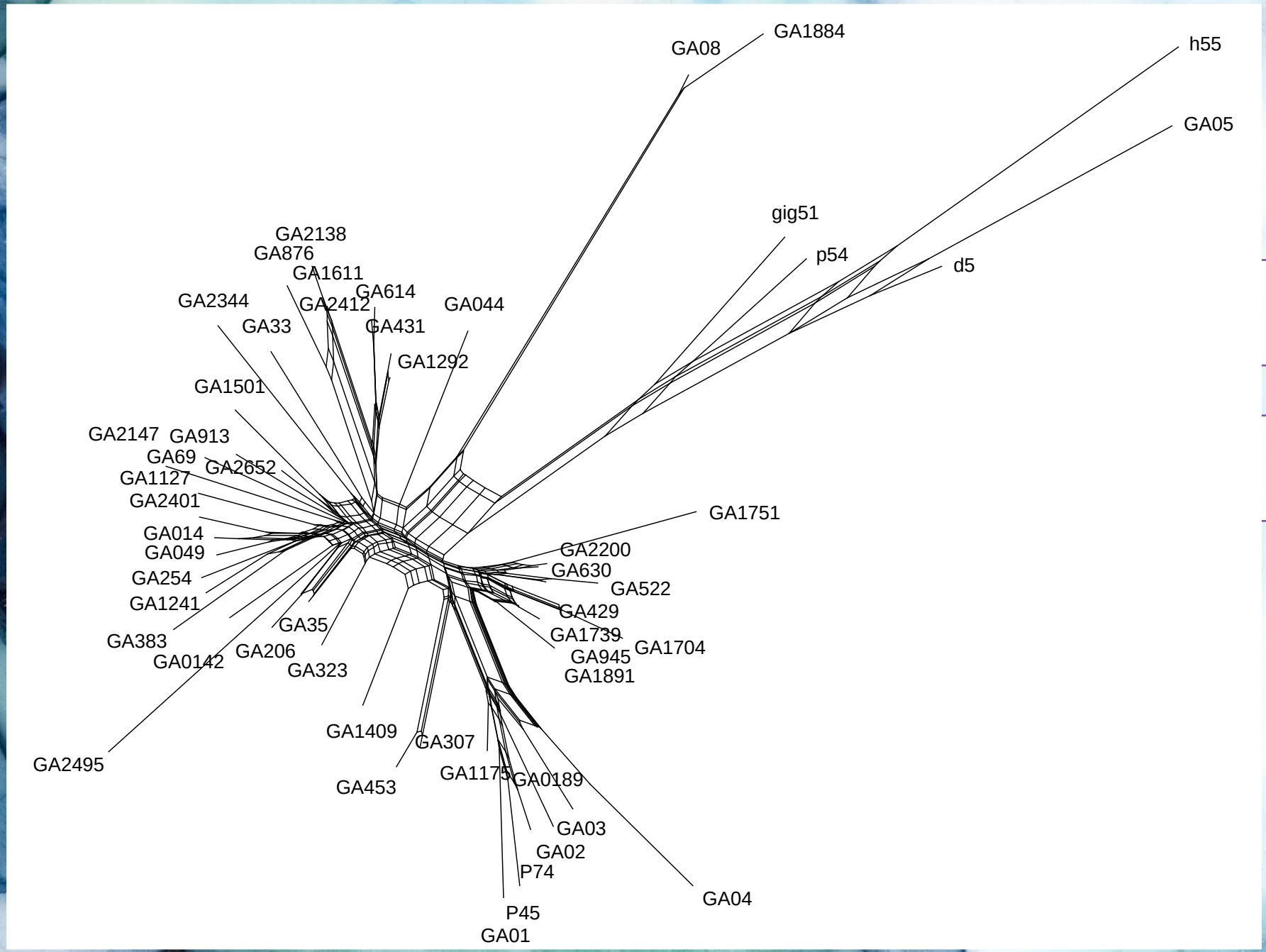
Prospects

Developing GUI for Relate…?

- It needs some studying to use Relate at this stage of development

Integrating tree inference and network methods

- Network analysis is promising


GA08
GA1884
h55
GA05
gig51
p54
d5
GA2138
GA876
GA1611
GA614
GA2344
GA2412
GA044
GA33
GA431
GA1292
GA1501
GA2147
GA913
GA69
GA2652
GA1127
GA2401
GA1751
GA014
GA049
GA2200
GA630
GA254
GA522
GA1241
GA429
GA35
GA1739
GA383
GA1704
GA945
GA0142
GA206
GA323
GA1891
GA1409
GA307
GA2495
GA1175
GA0189
GA453
GA03
GA02
P74
GA04
P45
GA01

Prospects

Developing GUI for Relate…?

- **It needs some studying to use Relate at this stage of development**

Integrating tree inference and network methods

- **Network analysis is promising**

Transcribing process must be automated

- **Using machine learning and neural networks**
  - Handwritten text recognition (HTR) techniques


Thank you!
Enjoy your time in Denver!