

Hands-on: Authorship Attribution Through Stylometry

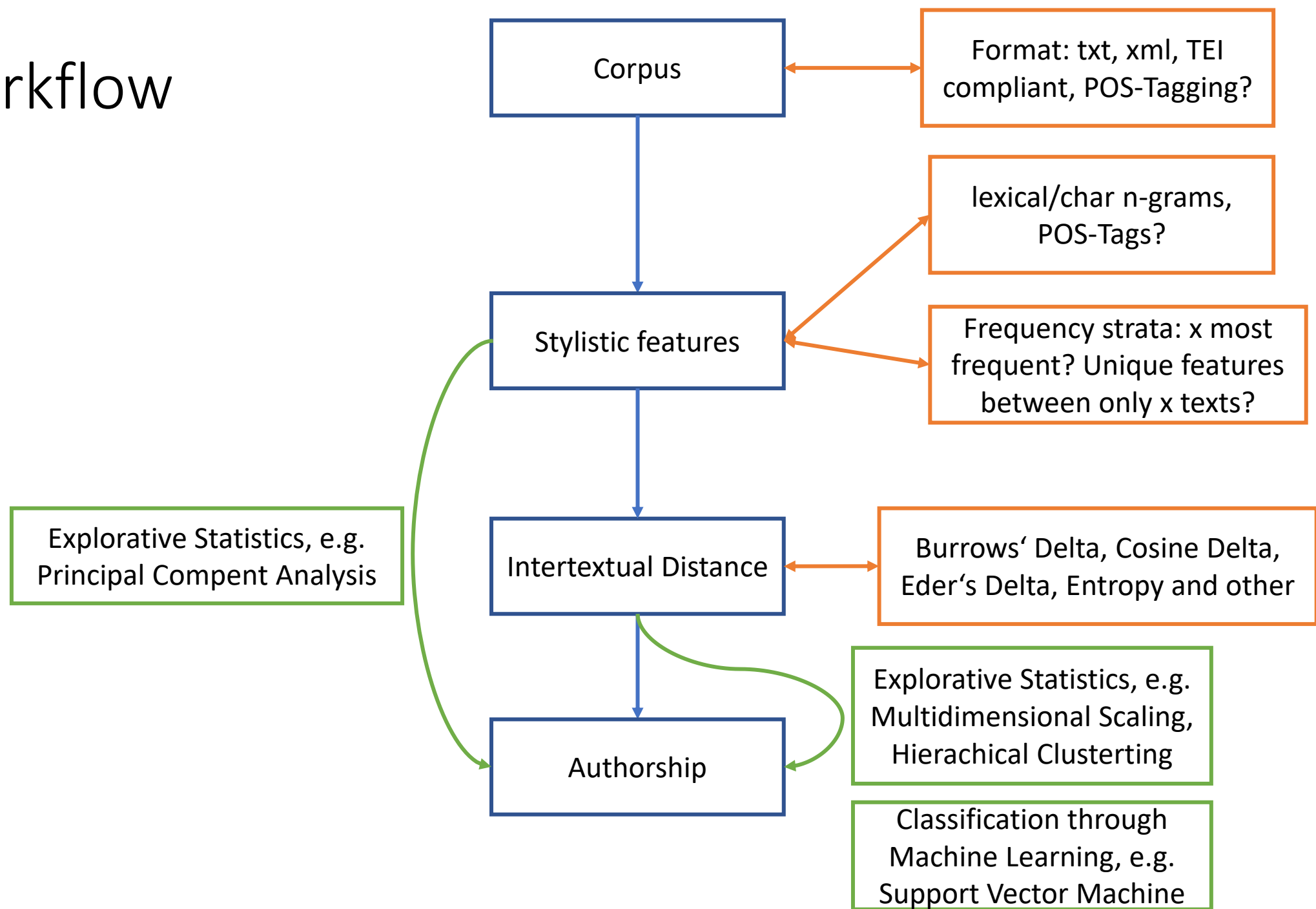
Nikola Krisztian Czindrity

ATDS 2022

Materials available at

<https://github.com/NKCZ/atds2022stylo>

Workflow



Good news: all covered in *Stylo*!

The following are required:

- R: <https://www.r-project.org/>
 - R for Windows: <https://cran.r-project.org/bin/windows/base/>
- RStudio: <https://www.rstudio.com/products/rstudio/>
- R-package *Stylo*:
 - run in your R-Enviroment: **install.packages("stylo")**
 - <https://cran.r-project.org/web/packages/stylo/index.html>
- OR
- try running it in Binder (see Github)

Corpus

Corpus

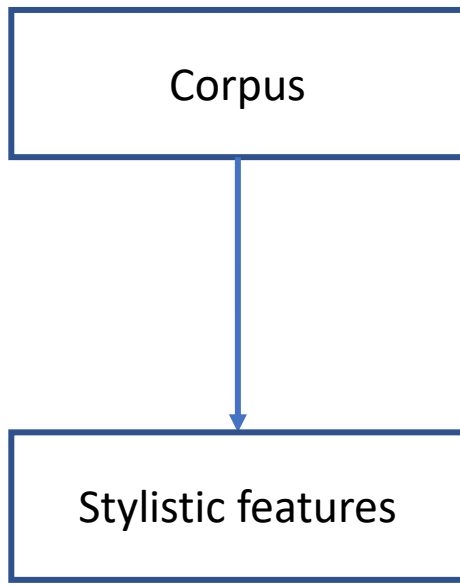
Format: txt, xml, TEI
compliant, POS-Tagging?



Corpus

Format: txt, xml, TEI
compliant, POS-Tagging?

Saga Corpus
48 XML (TEI)
Icelandic People Sagas,
Landn., Sturlunga s.



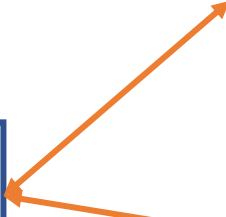
Saga Corpus
48 XML (TEI)
Icelandic People Sagas,
Landn., Sturlunga s.

Corpus



Stylistic features

lexical/char n-grams,
POS-Tags?



Frequency strata: x most
frequent? Unique features
between only x texts?

Saga Corpus
48 XML (TEI)
Icelandic People Sagas, Hmsk.
Landn., Sturlunga s.

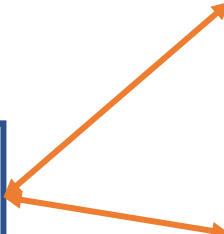
n-gram

- $S = \text{„Úlfur hét maður, son Bjálfa og Hallberu, dóttur Úlfs hins óarga.“}$
(*Egils saga*, Ch. 1)
- Lexical 1-gram:
 - $S_{1\text{-lex-gram}} = \{\text{„Úlfur“}, \text{„hét“}, \text{„maður“}, \text{„son“}, \text{„Bjálfa“}, \dots, \text{„óarga“}\}$
- Lexical 2-gram:
 - $S_{2\text{-lex-gram}} = \{\text{„Úlfur hét“}, \text{„hét maður“}, \text{„maður son“}, \text{„son Bjálfa“}, \text{„Bjálfa og“}, \dots, \text{„hins óarga“}\}$
- Character 4-gram:
 - $S_{4\text{-char_gram}} = \{\text{„Úlfu“}, \text{„lfur“}, \text{„fur “}, \text{„ur h“}, \dots, \text{„rga.“}\}$

Corpus



Stylistic features



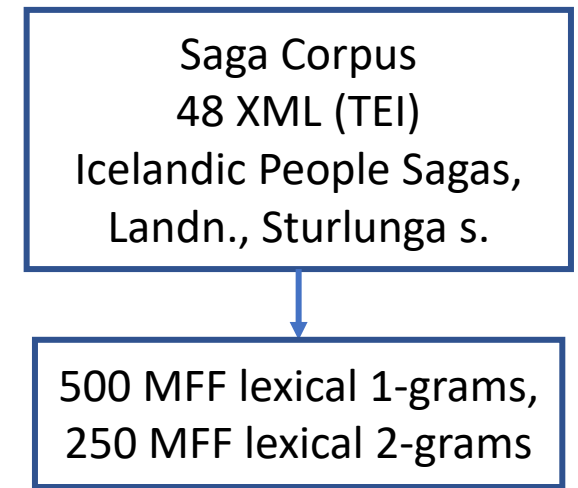
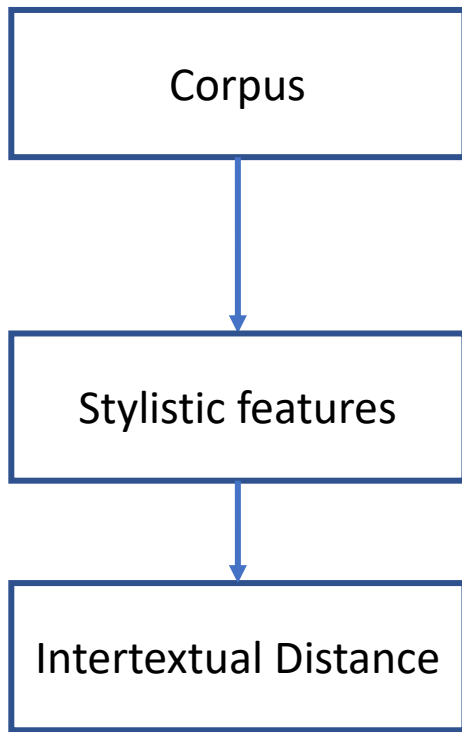
lexical/char n-grams,
POS-Tags?

Frequency strata: x most
frequent? Unique features
between only x texts?

Saga Corpus
48 XML (TEI)
Icelandic People Sagas,
Landn., Sturlunga s.



500 MFF lexical 1-grams,
250 MFF lexical 2-grams



Corpus



Stylistic features



Intertextual Distance



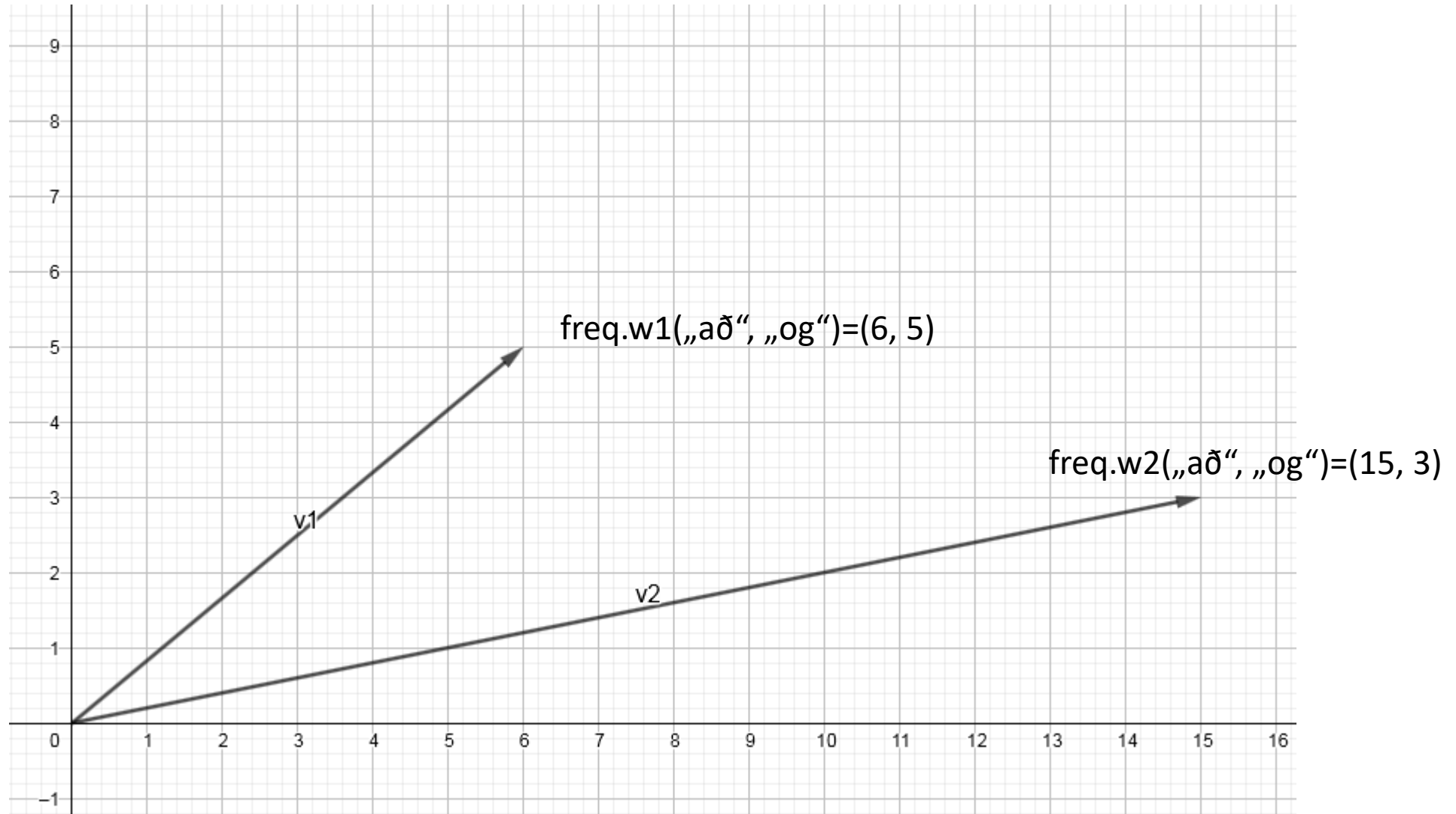
Burrows' Delta, Cosine Delta,
Eder's Delta, Entropy and other

Saga Corpus
48 XML (TEI)
Icelandic People Sagas,
Landn., Sturlunga s.

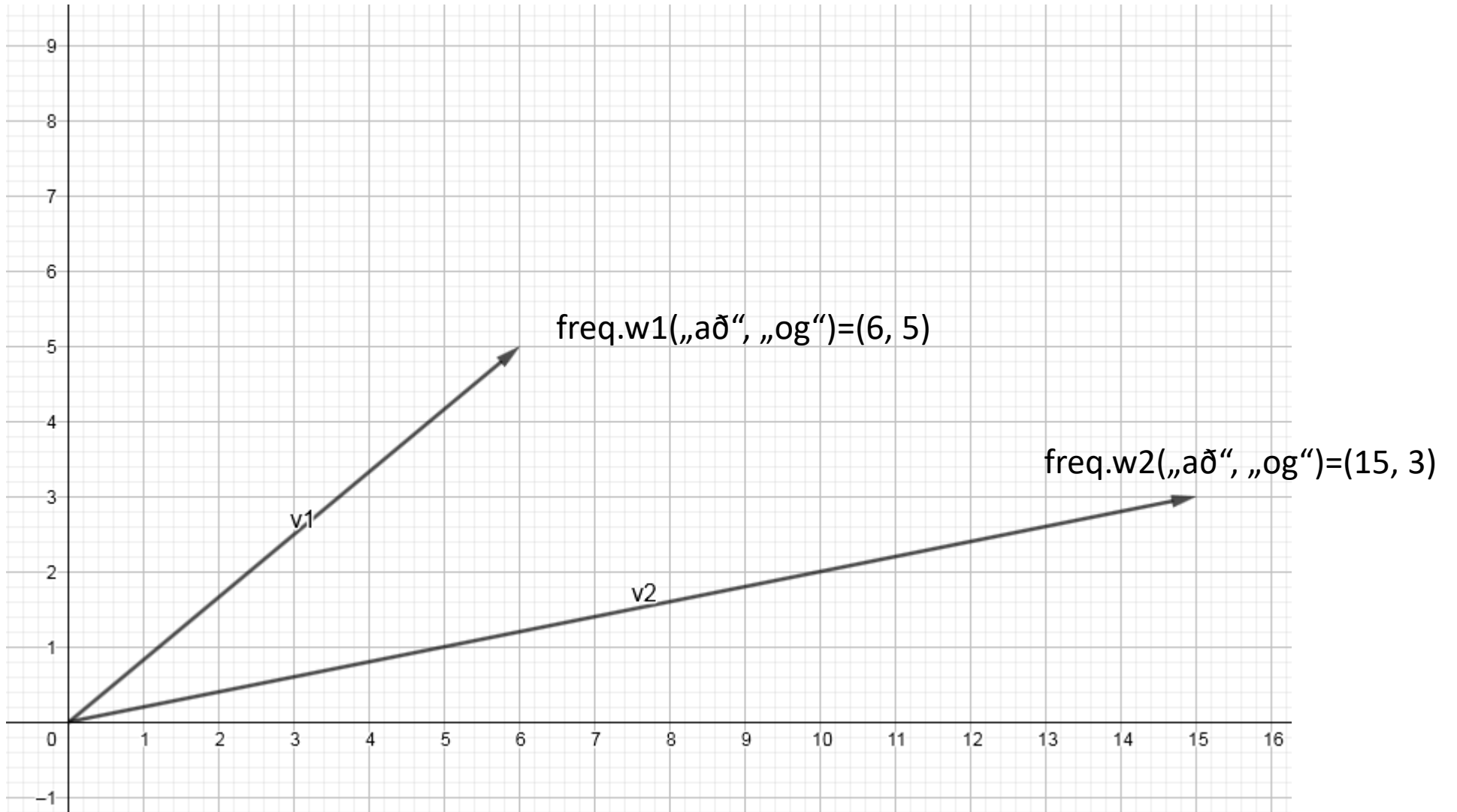


500 MFF lexical 1-grams,
250 MFF lexical 2-grams

Vector Space

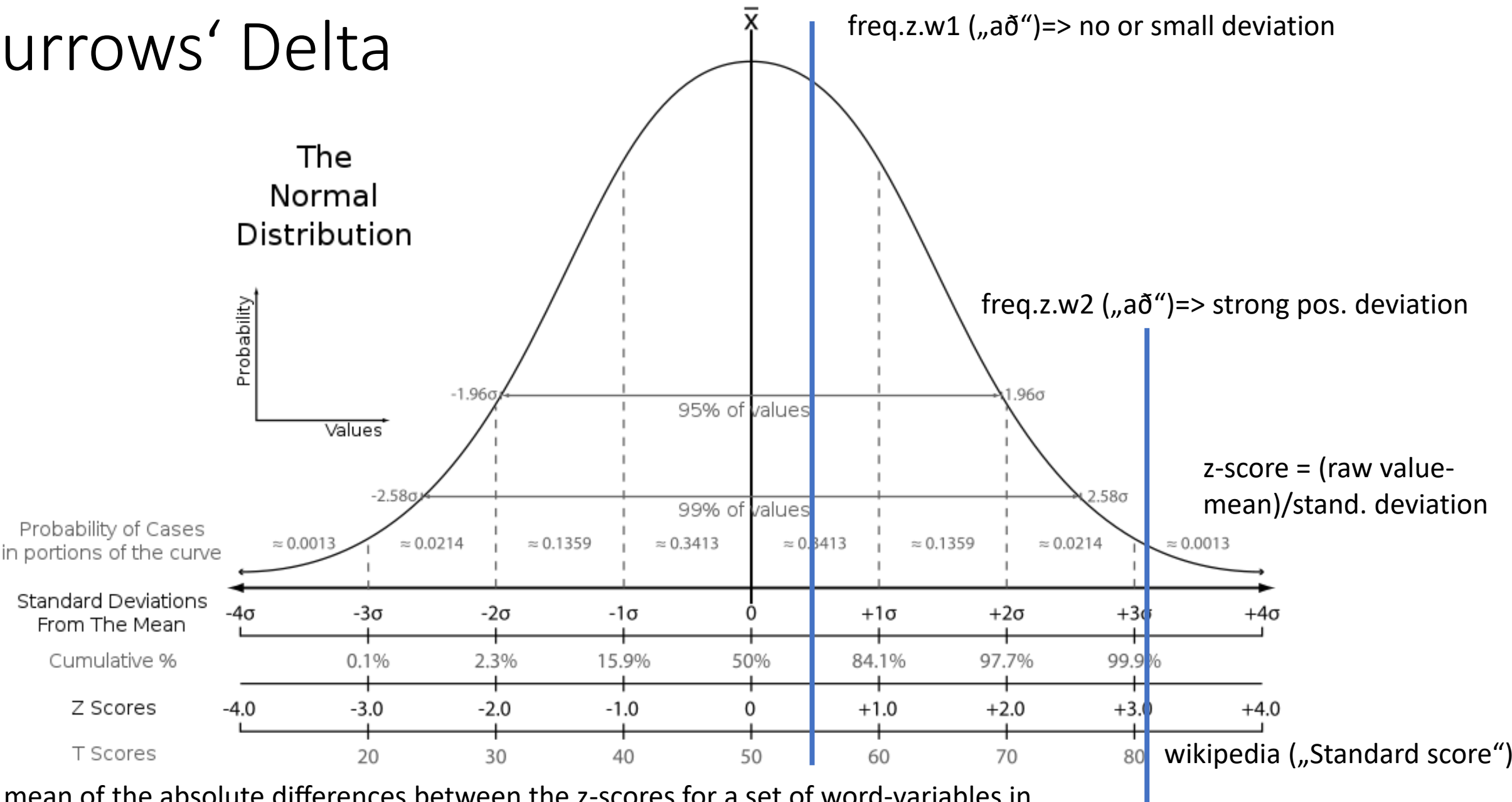


Vector Space



In that case 2-D, but it can be generalized to n-Dimensions.

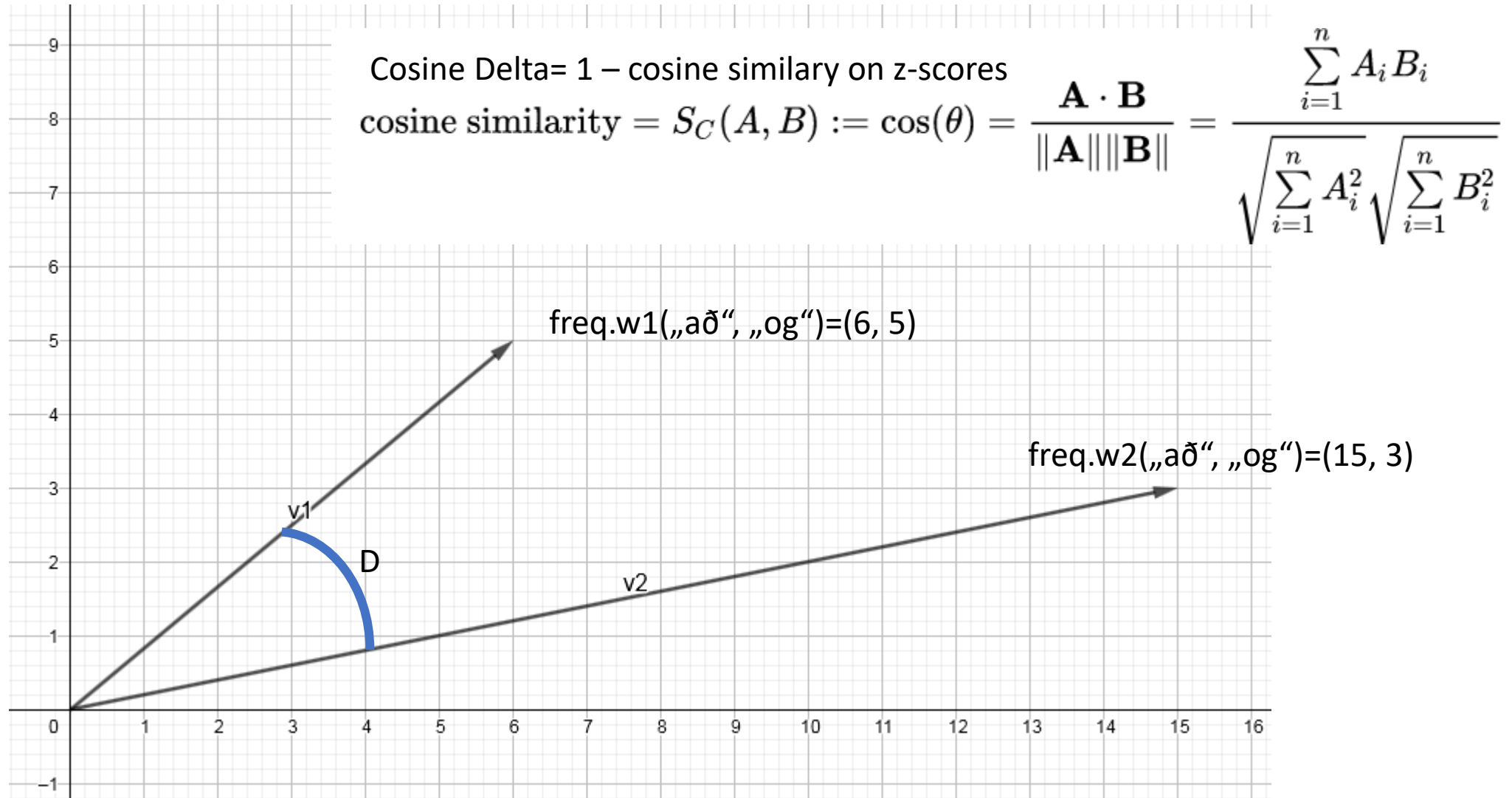
Burrows' Delta



“the mean of the absolute differences between the z-scores for a set of word-variables in a given text-group and the z-scores for the same set of word-variables in a target text.” Burrows 2002

Cosine Similarity & Delta

Wikipedia („cosine similarity“)



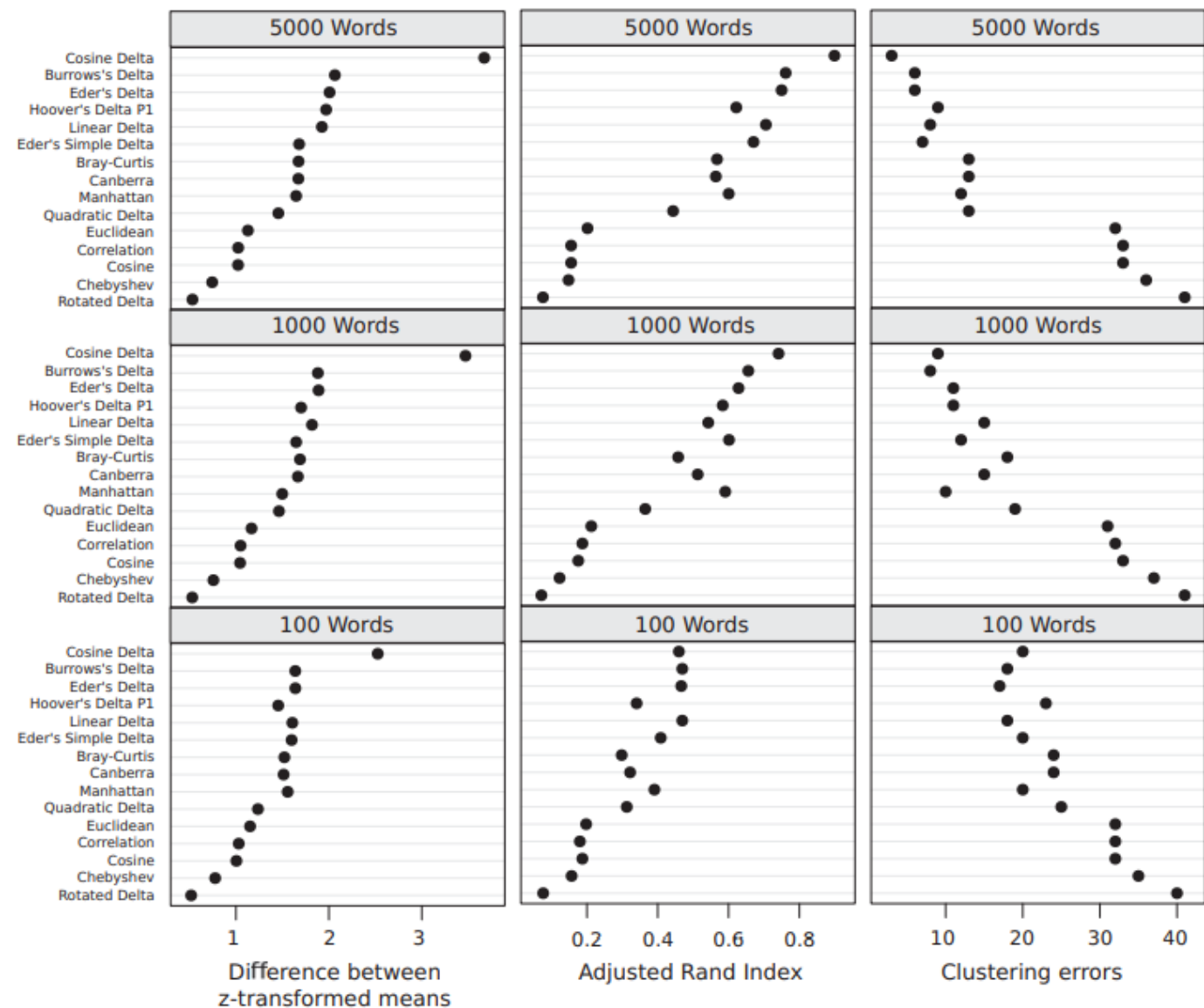


Fig. 4 Performance of distance measures on English texts. Indicated in terms of the difference between z-transformed means of ingroup (same author) and outgroup distances (different authors), as Adjusted Rand Index (higher values indicate better differentiation), and in terms of clustering errors (lower values indicate better differentiation). Distance measures are sorted according to their maximum performance in all test conditions. The non-Delta measures are popular basic distance measures on raw relative frequencies. Similar results for French and German

Corpus



Stylistic features



Intertextual Distance



Burrows' Delta, Cosine Delta,
Eder's Delta, Entropy and other

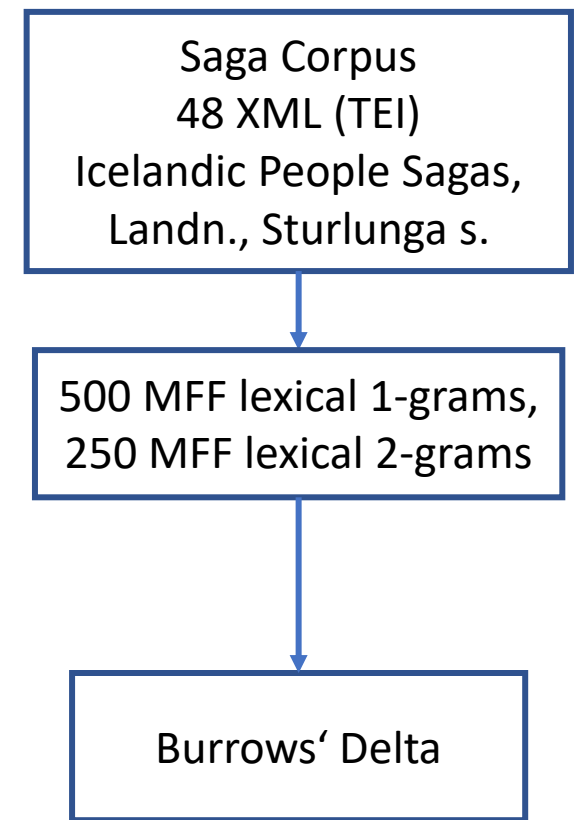
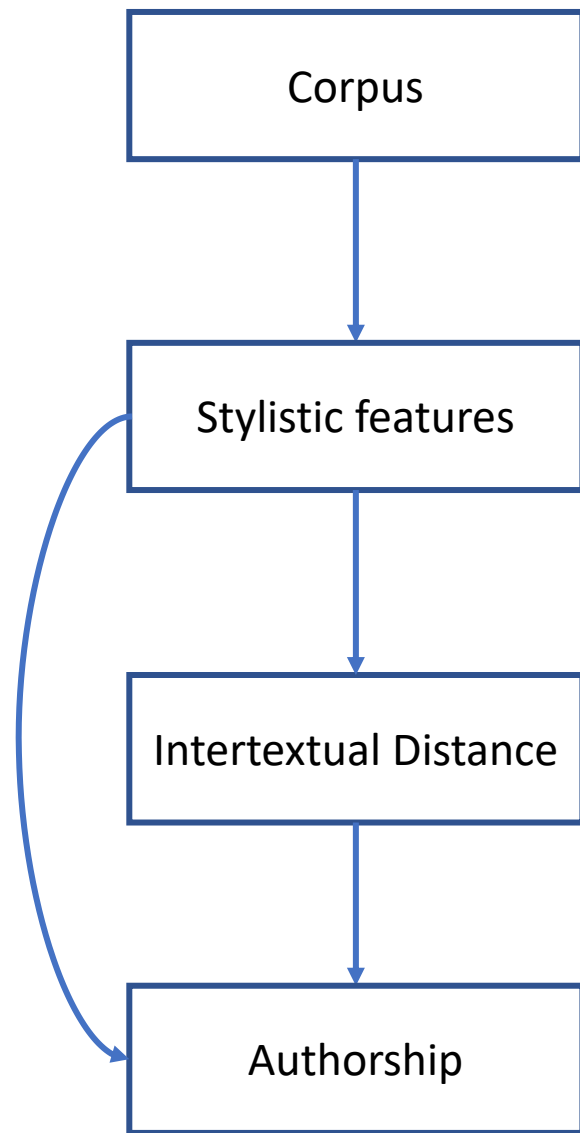
Saga Corpus
48 XML (TEI)
Icelandic People Sagas,
Landn., Sturlunga s.

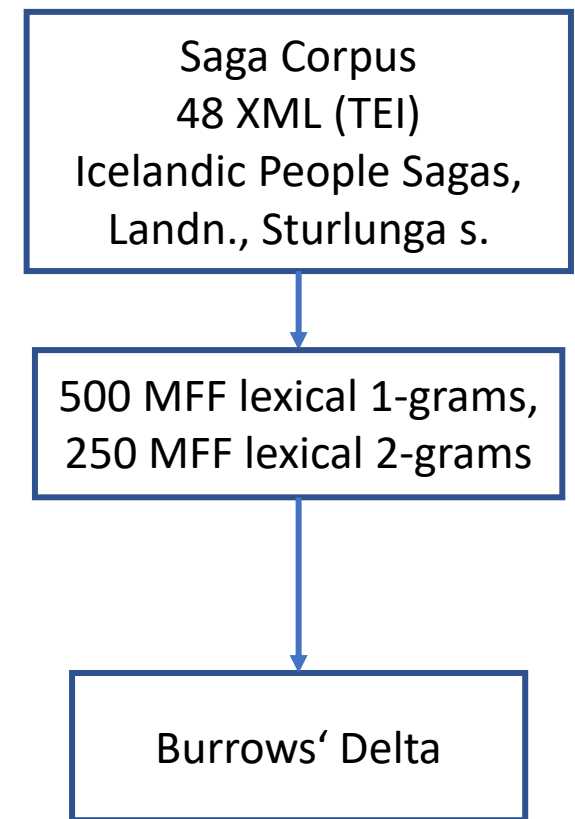
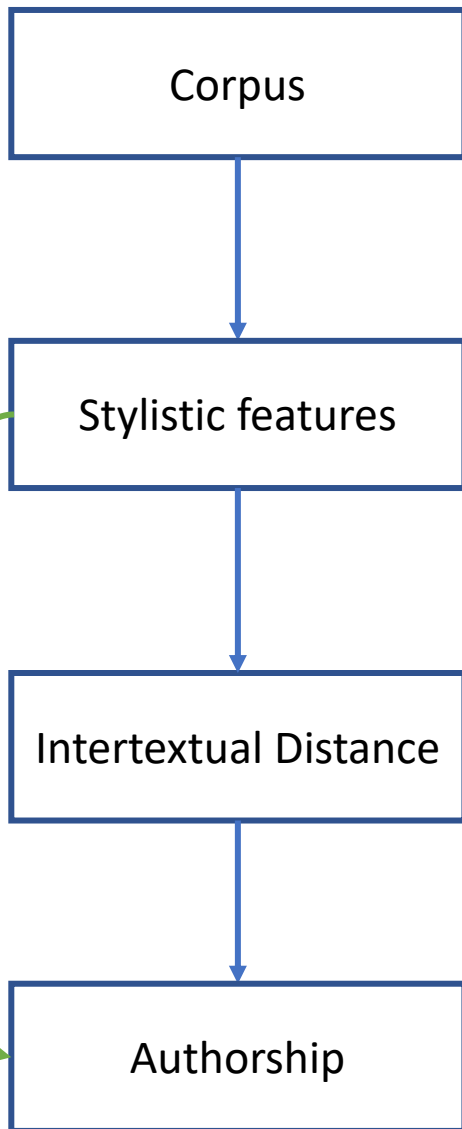


500 MFF lexical 1-grams,
250 MFF lexical 2-grams



Burrows' Delta





Corpus



Stylistic features



Intertextual Distance



Authorship

Explorative
Statistics,
e.g. PCA



Explorative Statistics, e.g.
MDS, HCA



Saga Corpus
48 XML (TEI)
Icelandic People Sagas,
Landn., Sturlunga s.



500 MFF lexical 1-grams,
250 MFF lexical 2-grams



Burrows' Delta

Corpus



Stylistic features



Intertextual Distance

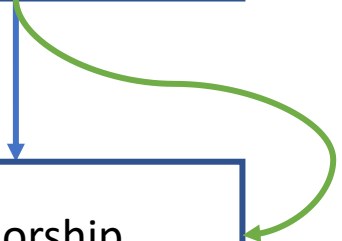


Authorship

Explorative Statistics,
e.g. PCA



Explorative Statistics, e.g.
MDS, HCA



Classification through
Machine Learning, e.g.
Support Vector Machine

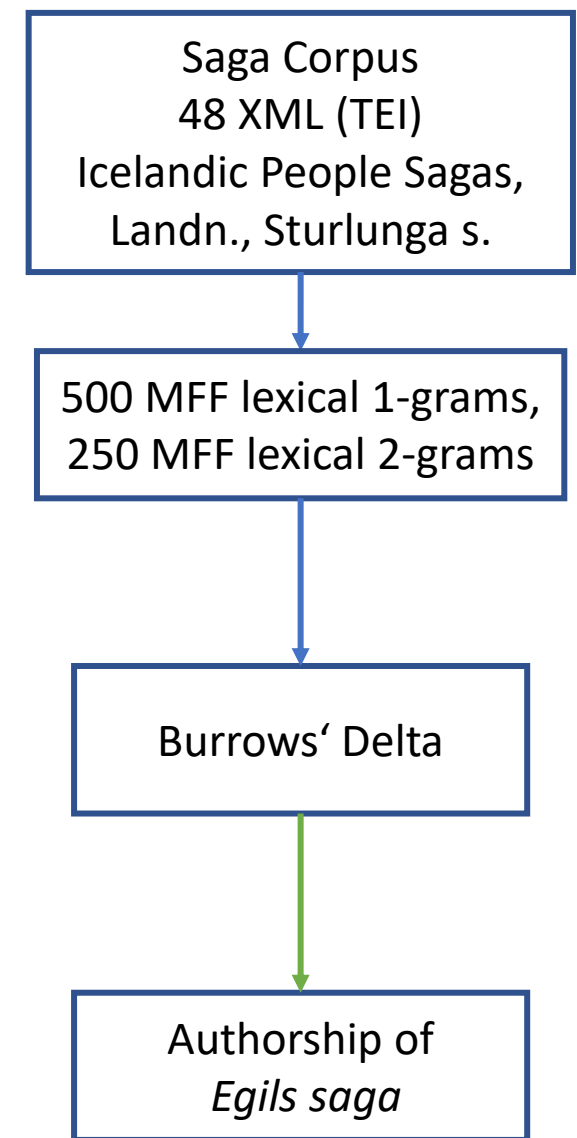
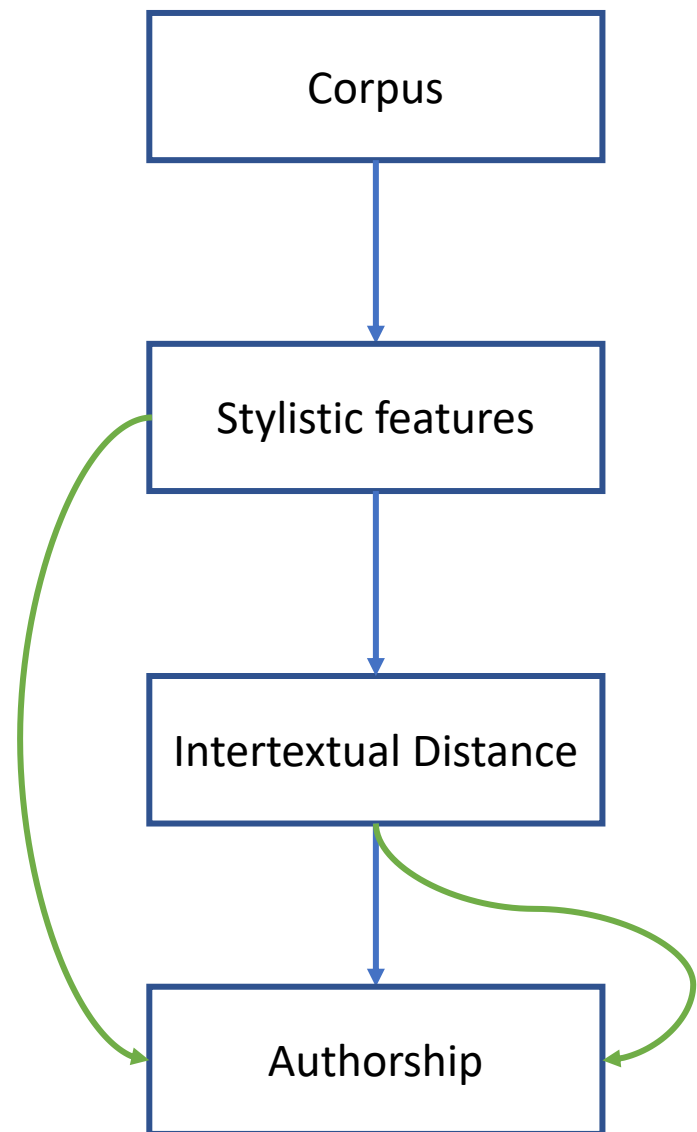
Saga Corpus
48 XML (TEI)
Icelandic People Sagas,
Landn., Sturlunga s.

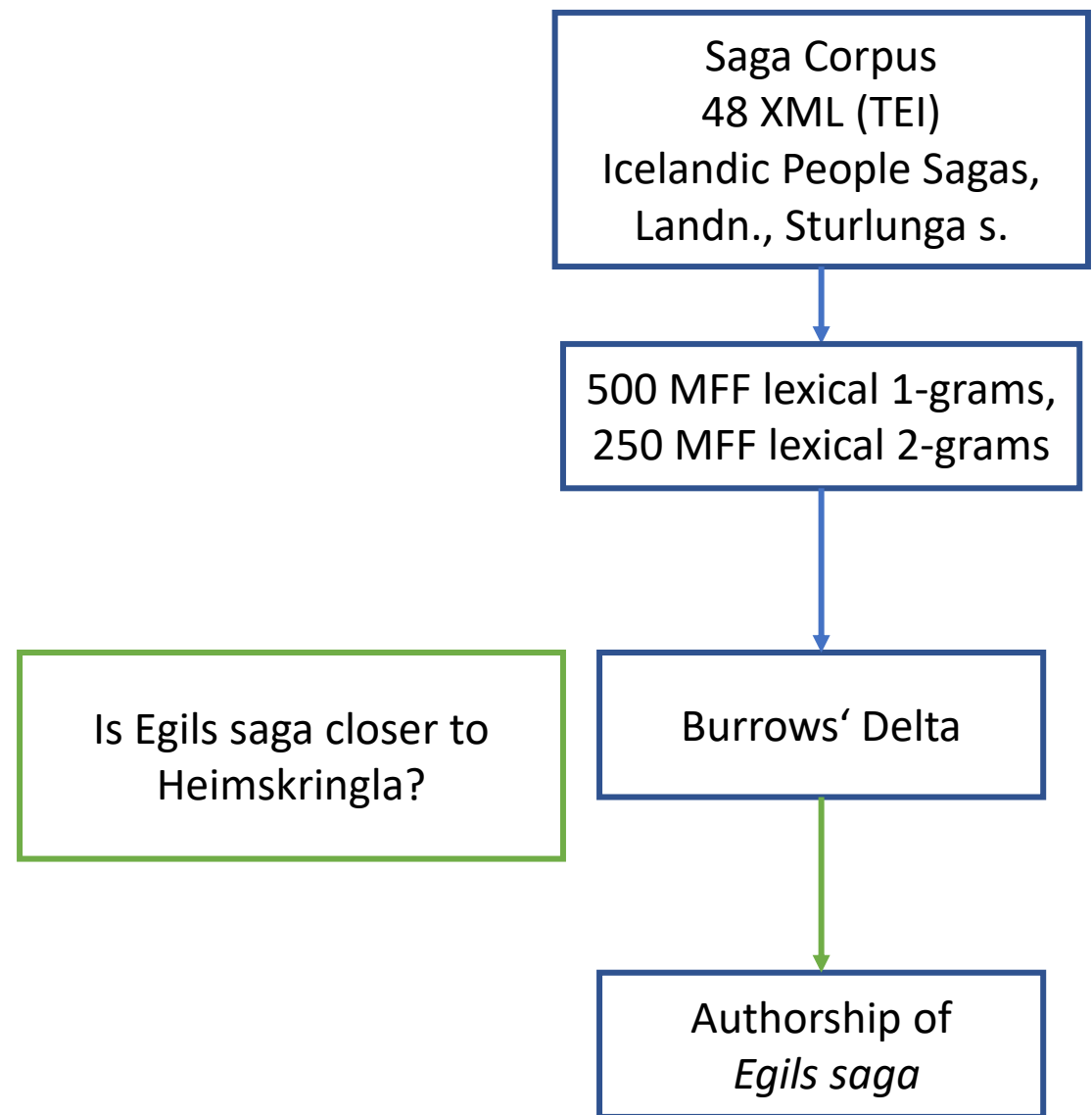
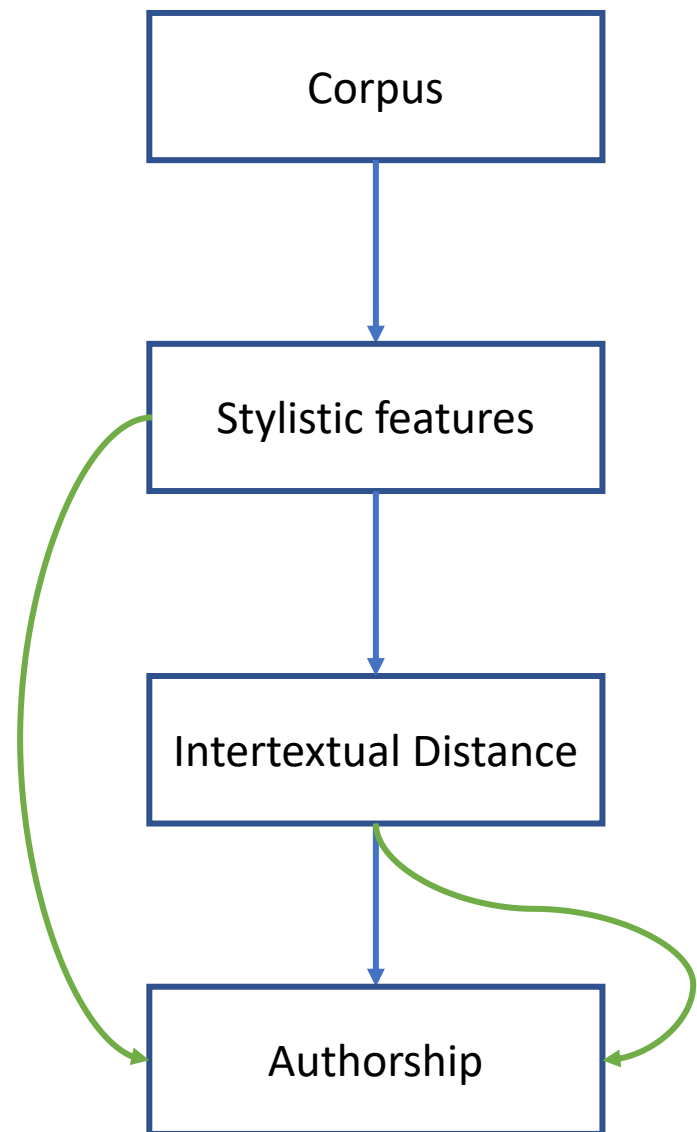


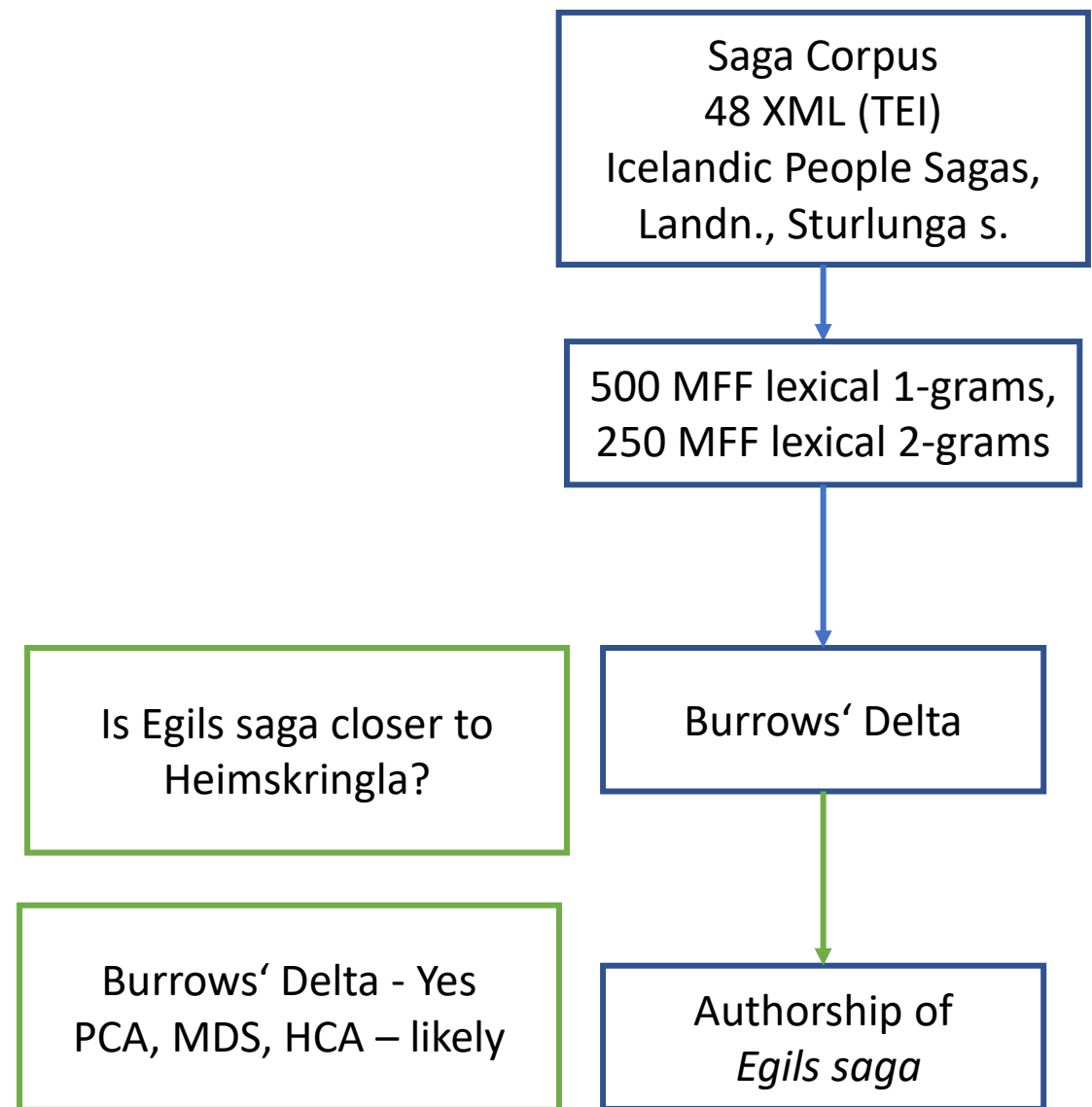
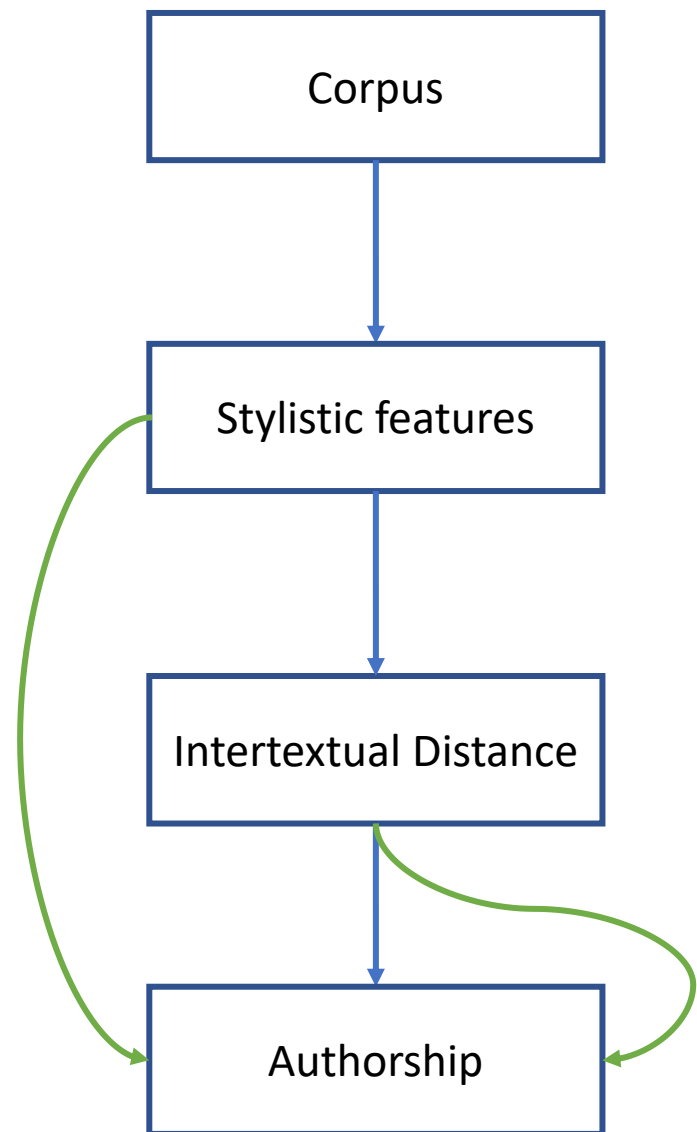
500 MFF lexical 1-grams,
250 MFF lexical 2-grams

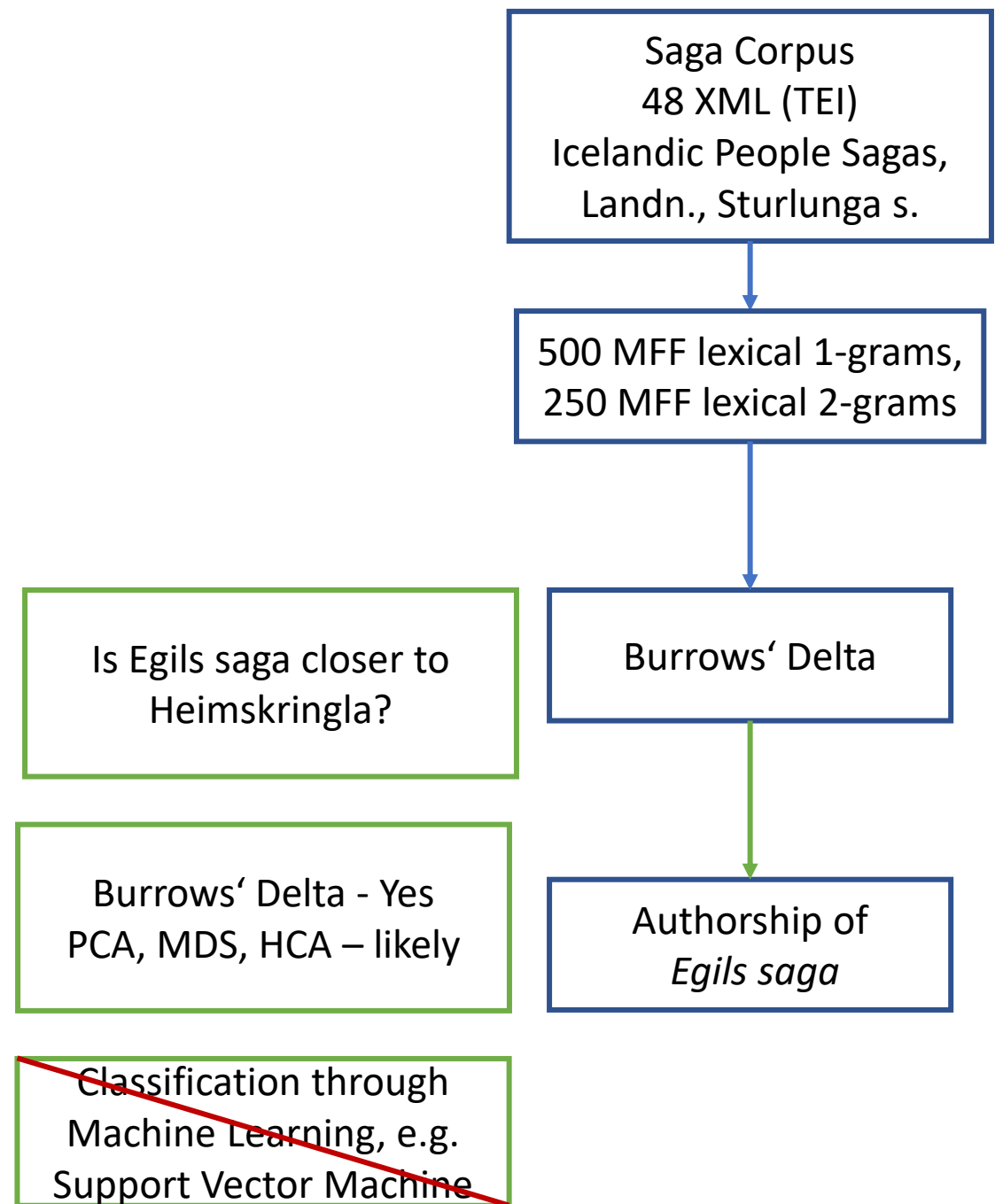
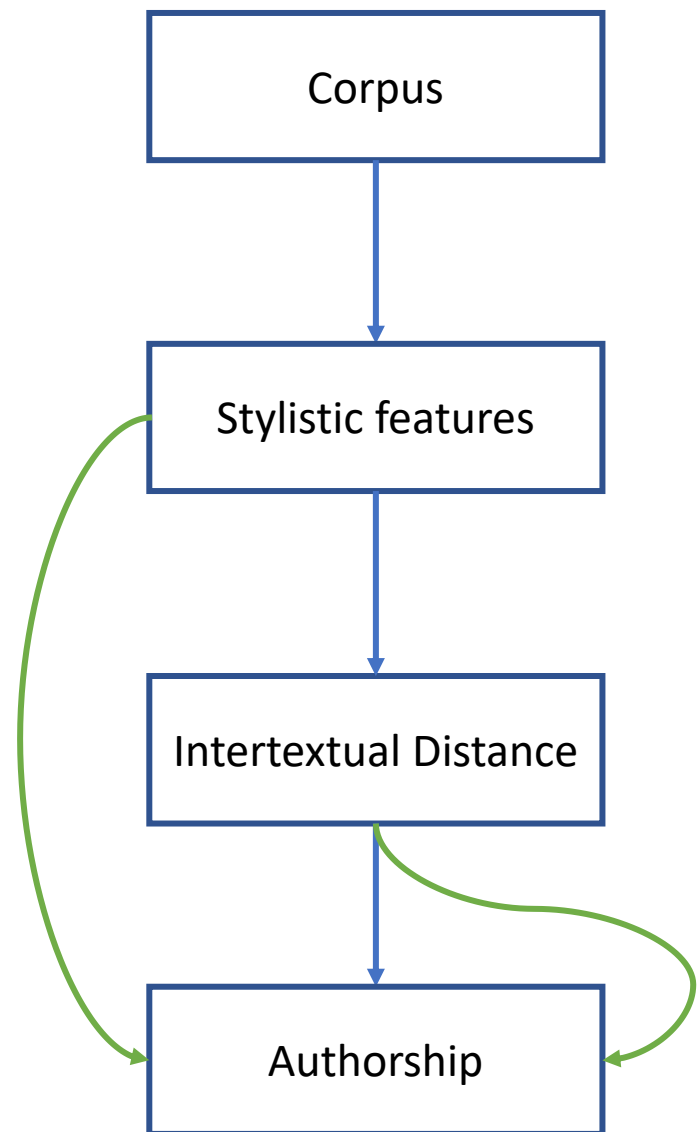


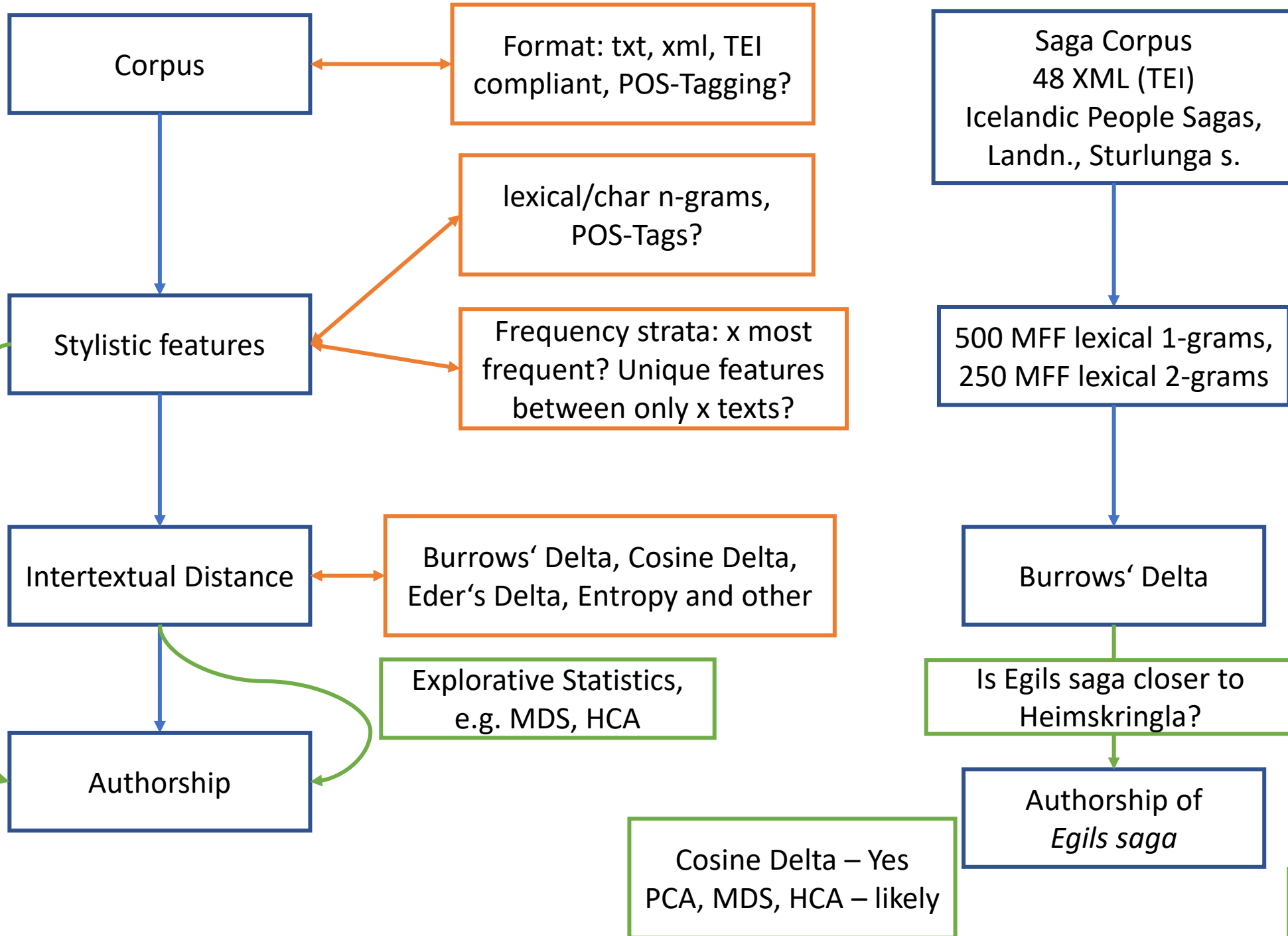
Burrows' Delta











Explorative Statistics, e.g. PCA

Explorative Statistics, e.g. MDS, HCA

Cosine Delta – Yes
PCA, MDS, HCA – likely

~~ML~~

Sources & further Information I

- Materials at <https://github.com/NKCZ/atds2022style>
- Corpus:
 - *Saga Corpus* = The Saga Corpus, ed. Eiríkur Rögnvaldsson / Sigrún Helgadóttir. (online: URL <http://www.malfong.is/index.php?lang=en&pg=fornritin>; 2021).
- General information on stylometry:
 - Patrick Juola, Authorship Attribution. In: Foundations and Trends in Information Retrieval 1(3) (2007), 233-334.
 - Efstathios Stamatatos, A Survey of Modern Authorship Attribution Method. In: JASIST 60(3) (2008), 539-556.
 - Maciej Eder / Jan Rybicki / Mike Kestemon, Stylometry with R: a package for computational text analysis. In: R Journal 8(1) (2016), 107-121.
 - Stefan Evert / Thomas Proisl / Fotis Jannidis / Isabella Reger / Steffen Pielström / Christof Schöch / Thorsten Vitt, Understanding and explaining Delta measures for authorship attribution. In: DSH 32(Suppl. 2) (2017), 4–16.
 - Burrows 2002 = John Burrows, 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. In: LLC 17(3) (2002), 267-287.
 - Burrows 2007 = John Burrows, All the Way Through: Testing for Authorship in Different Frequency Strata. In: LLC 22(1) (2007), 27-47.

Sources & further Information II

- Authorship through stylometry: *Egils saga*
 - Sigurður Ingibergur Björnsson / Steingrímur Páll Kárason / Jón Karl Helgason, Stylometry and the Faded Fingerprints of Saga Authors. In: In Search of the Culprit. Aspects of Medieval Authorship, ed. Lukas Rösli / Stefanie Gropper (Berlin/Boston 2021), 97-122.
 - Haukur Þorgeirsson, How similar are Heimskringla and Egils saga? An application of Burrows' delta to Icelandic texts. In: EJSS 48(1) (2018), 1-18.