How Do You Measure Style?

(and much more)

Misha Sonkin

Preliminary Information

- **Primary goal:** tell you about Burrows' Delta and look "under the hood".
- **Secondary goal:** tell you about the project.
- Workshop-Talk Hybrid.

Overview

- Introduction to Stylometry
- Delta
  - Method
  - Authorship Attribution
  - Translator Comparison
  - Boris Pasternak (My Study)
- Discussion

What is stylometry?

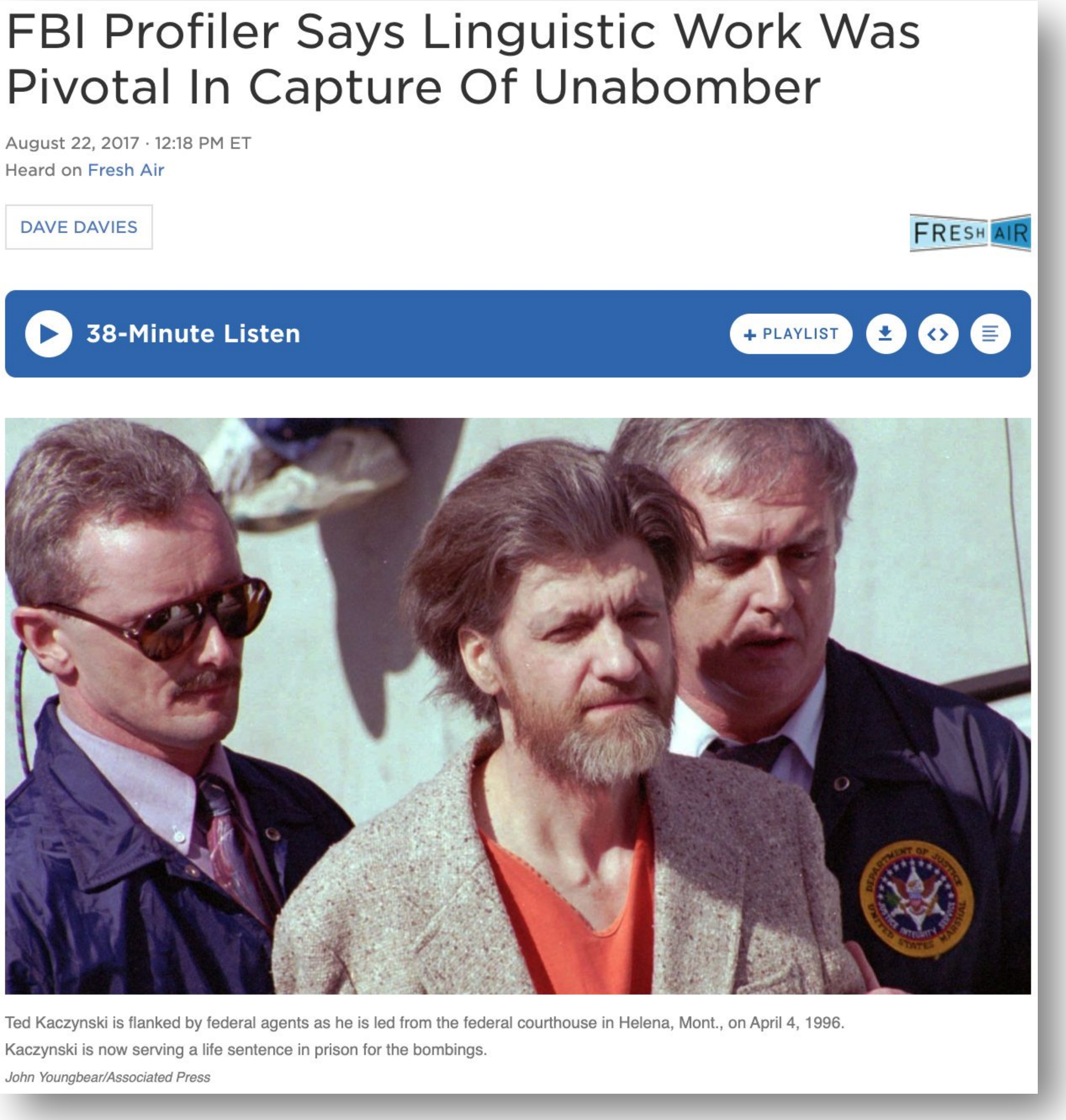
- Application of statistical methods in the study of **style**.
- Not just writing style! (Liu et al. 2016)

(a) f249 (b) f371 (c) f522 (d) f752

Why? 🤷

- Authorship Attribution
- Style Comparison
  - Authors
  - Translators
  - Human- vs. Machine-generated text
  - etc

Why? 🤷



FBI Profiler Says Linguistic Work Was Pivotal In Capture Of Unabomber

August 22, 2017 · 12:18 PM ET
Heard on Fresh Air

DAVE DAVIES

Ted Kaczynski is flanked by federal agents as he is led from the federal courthouse in Helena, Mont., on April 4, 1996. Kaczynski is now serving a life sentence in prison for the bombings.

John Youngbear/Associated Press

Why? 🤷

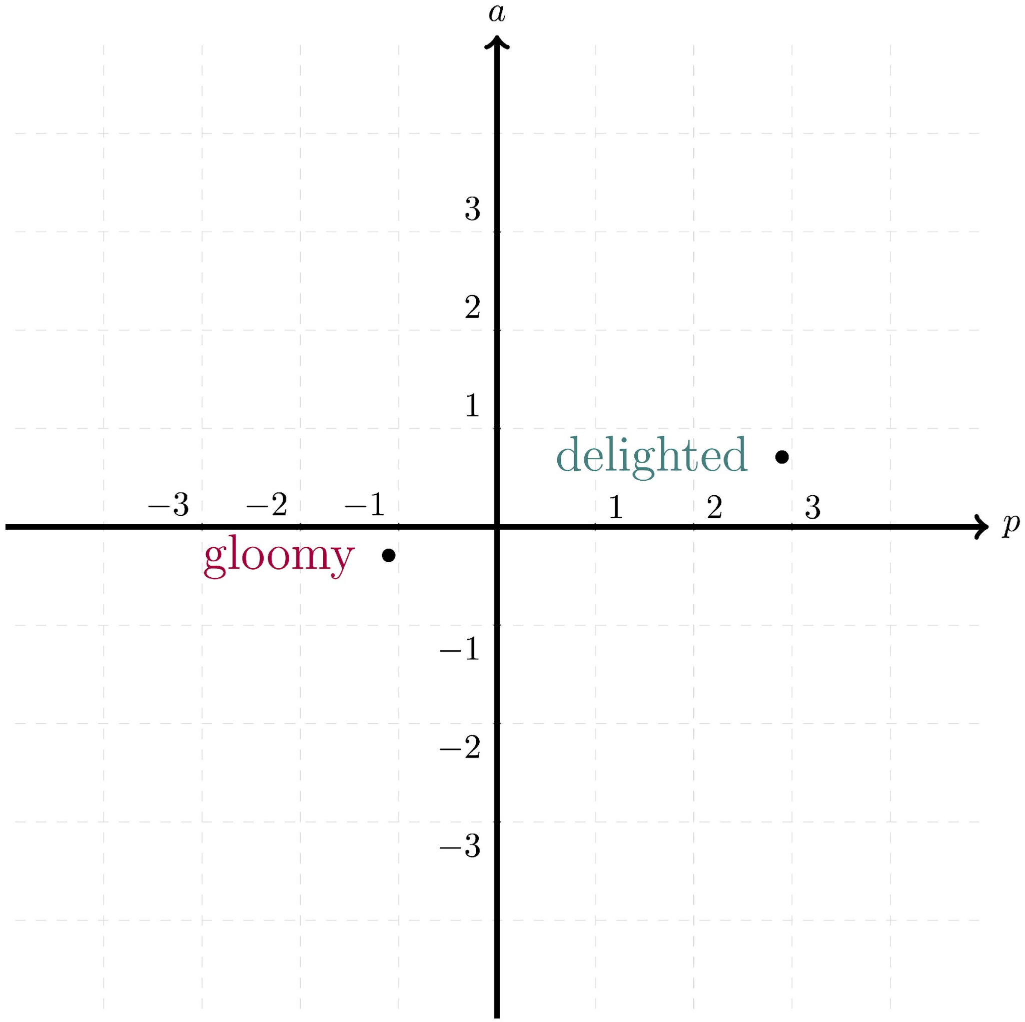
- Authorship Attribution
- Style Comparison
  - Authors
  - Translators
  - Human- vs. Machine-generated text
  - etc...
- Catching the Unabomber

Emotionality of The Beatles (Whissell 1996)

- Dictionary of Affect: crowd-sourced data on English words.
- Two measures, 7-point scale: *pleasantness* and *activation*.

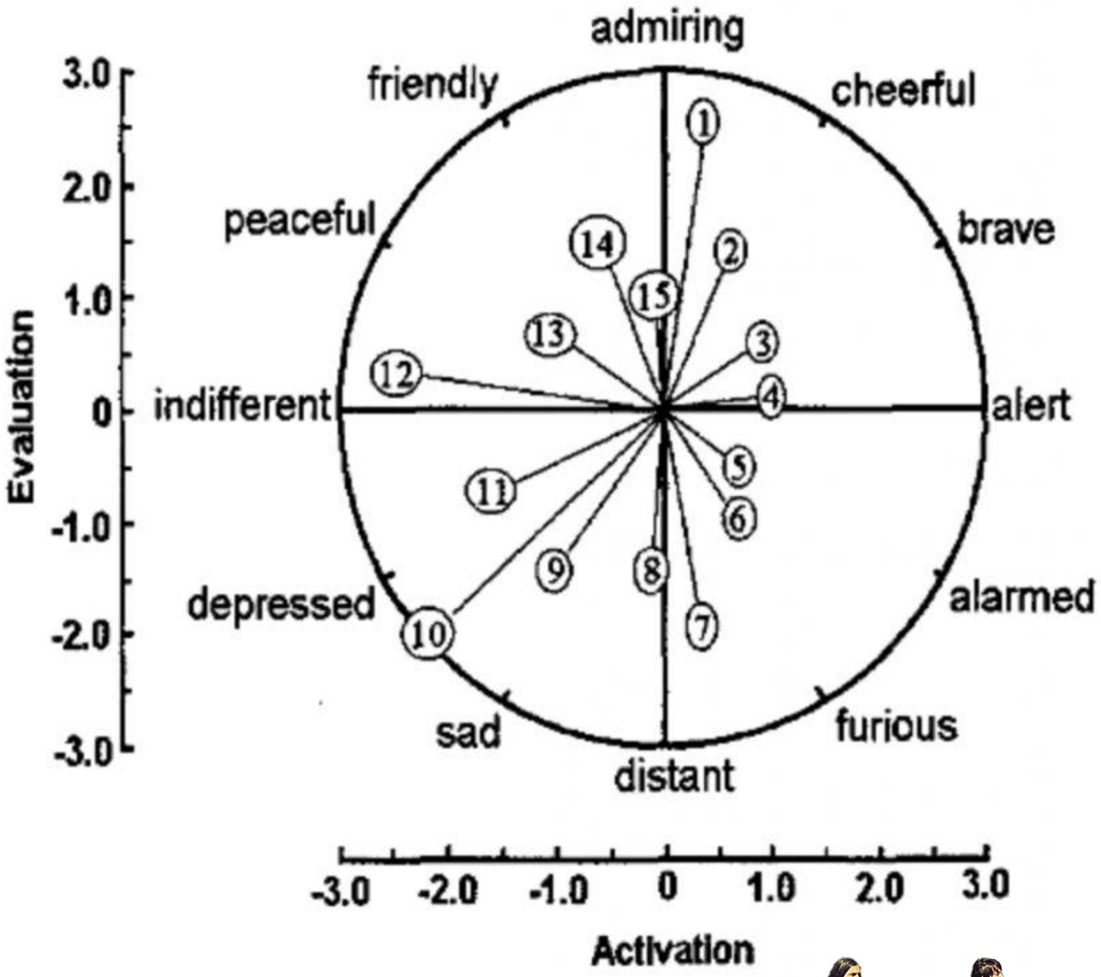
	"gloomy"	"delighted"
P	2.4	6.4
A	3.2	4.2

Emotionality of The Beatles (Whissell 1996)



Emotionality of The Beatles

- Used these measure to compare Paul McCartney’s song to John Lennon’s.
- Found some differences (Paul uses more “happy” words, etc.)
- **Problems**:
  - Specific only to the **English** language
  - ...and a very specific **database** (with dubious origins...)

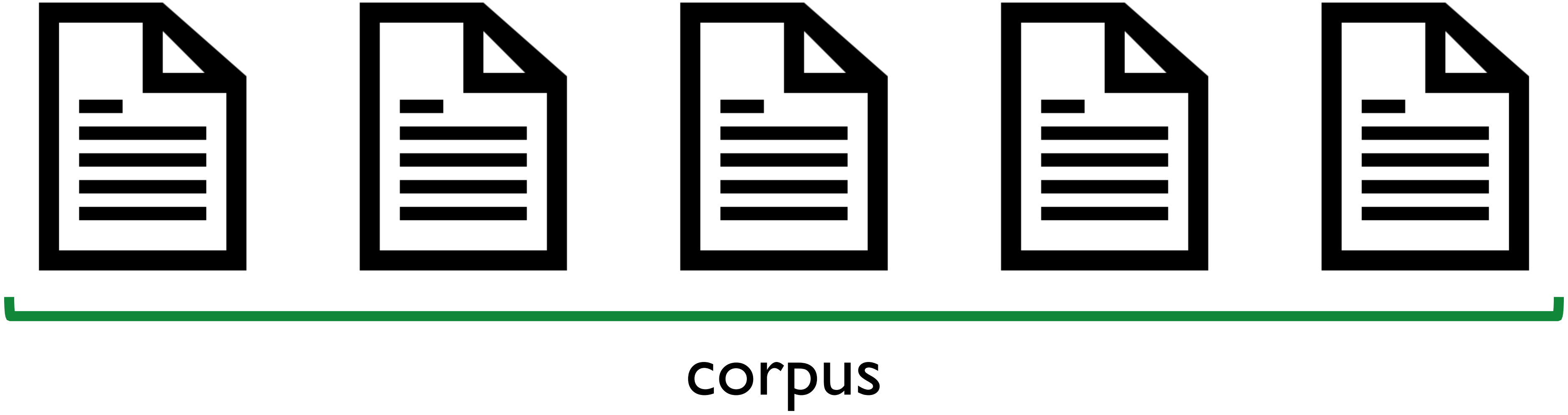


Is there a way to take language out of the picture?

Statistical measures?

Burrows’ Delta (Burrows 2002)

Burrows’ Delta (Burrows 2002)



Burrows' Delta (Burrows 2002)

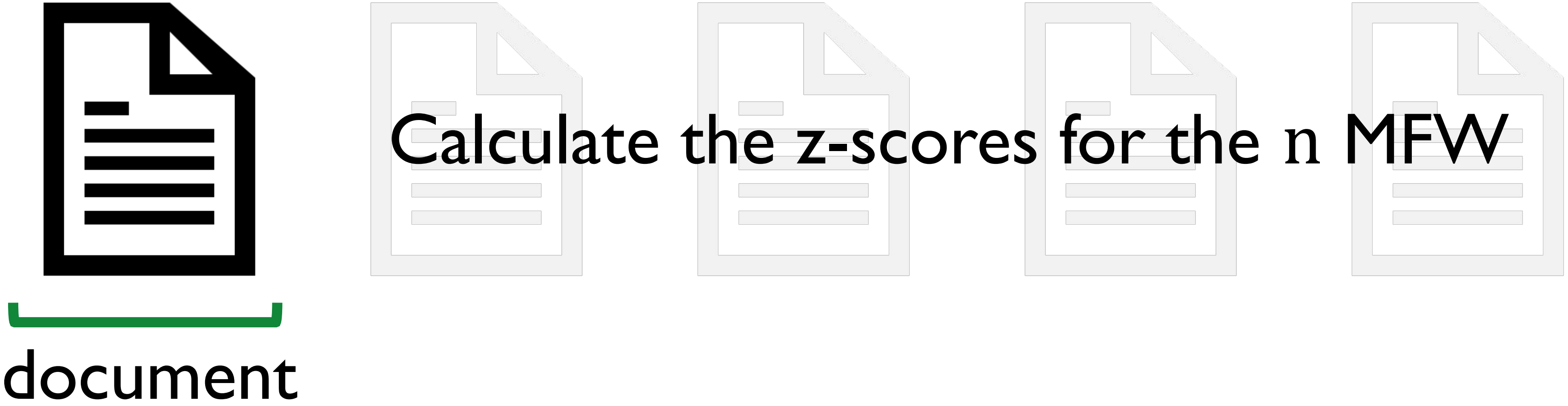
corpus

Calculate n **most frequent words** (MFW)
across the corpus

Burrows’ Delta (Burrows 2002)

document

Burrows' Delta (Burrows 2002)



Calculate the z-scores for the n MFW

document

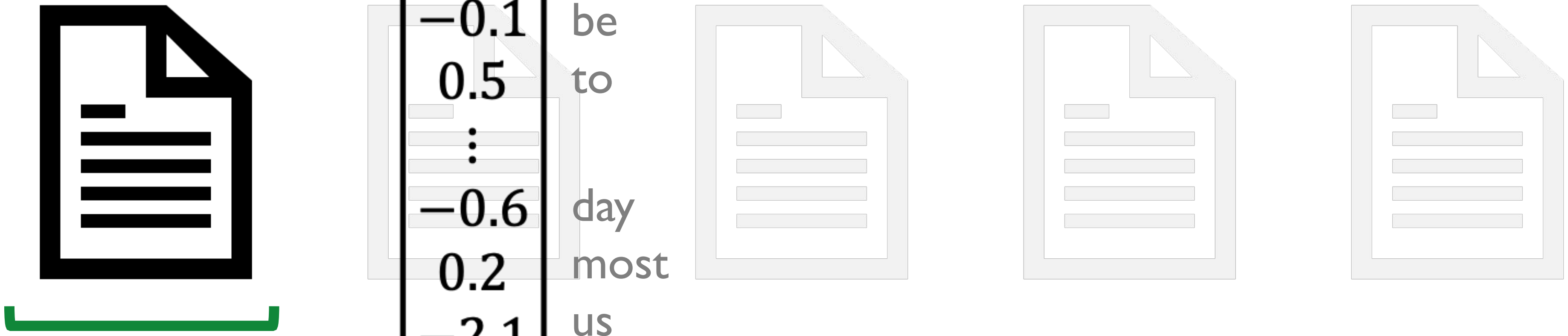
z-score

- A way no normalize the frequency.
- A z-score of word i in document D:

$$z_i(D) = \frac{D_i - \mu_i}{\sigma_i}$$

- Where
  - D_i– word frequency (in the document)
  - μ_i – mean of word’s frequency (in the corpus)
  - σ_i – standard deviation of word’s frequency (in the corpus)

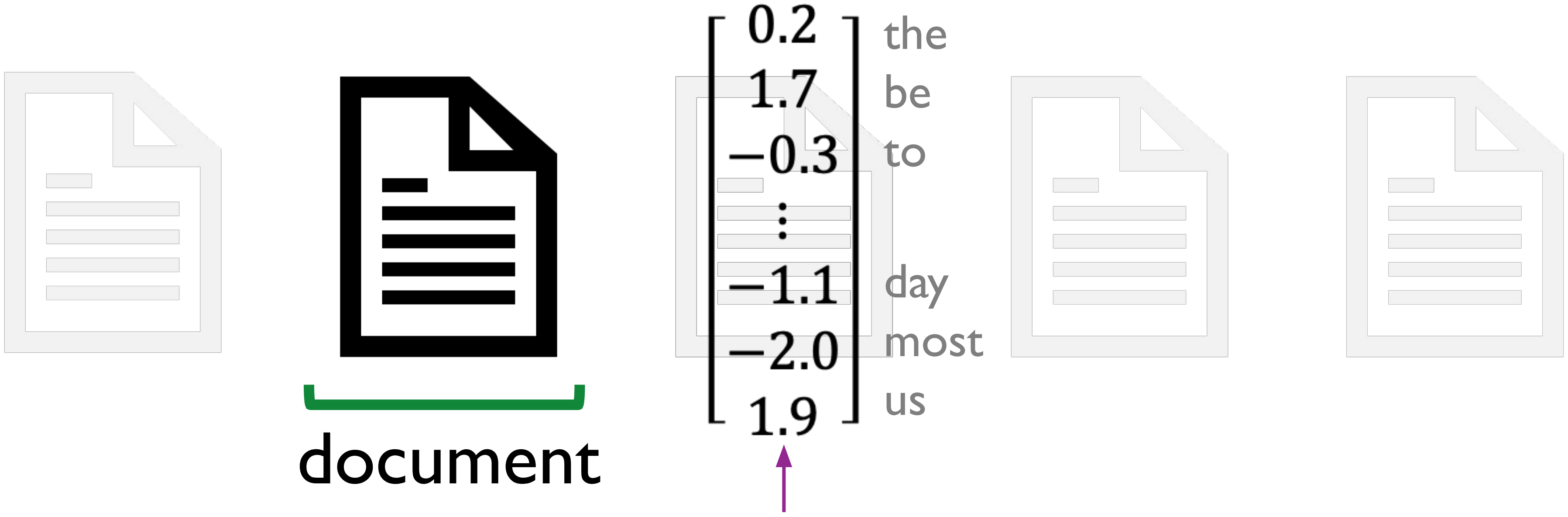
Burrows’ Delta (Burrows 2002)



$$\begin{array}{rl} \begin{bmatrix} 1.2 \\ -0.1 \\ 0.5 \\ \vdots \\ -0.6 \\ 0.2 \\ -2.1 \end{bmatrix} & \begin{array}{l} \text{the} \\ \text{be} \\ \text{to} \\ \\ \text{day} \\ \text{most} \\ \text{us} \end{array} \end{array}$$

document

Burrows' Delta (Burrows 2002)



$$\left[\begin{array}{c}0.2\\1.7\\-0.3\\\vdots\\-1.1\\-2.0\\1.9\end{array}\right]\begin{array}{l}\text{the}\\\text{be}\\\text{to}\\\\\text{day}\\\text{most}\\\text{us}\end{array}$$

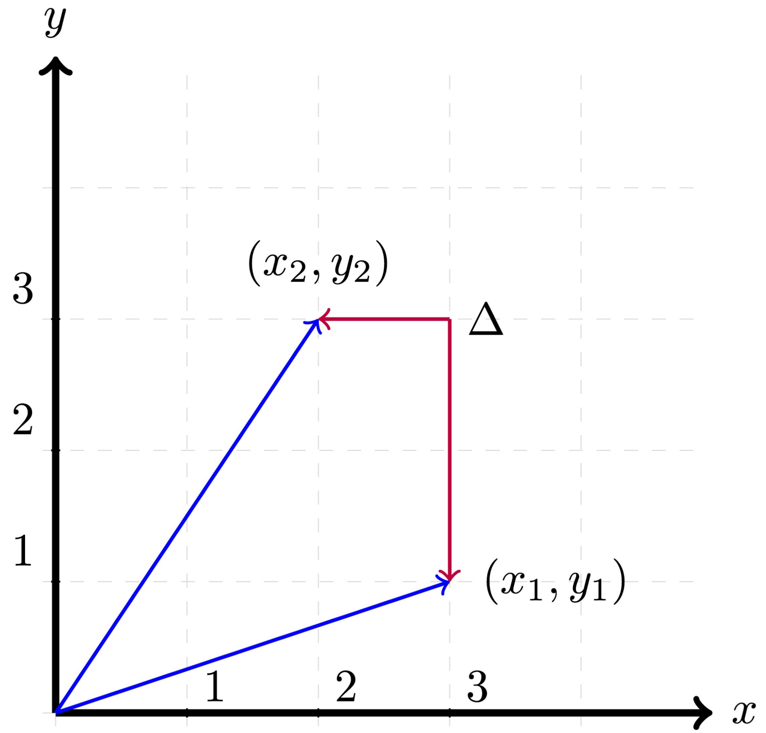
document

same words,
but different numbers!

Wait, are those...

- Yup, these are vectors!
- You can calculate the distance between two vectors (=documents)!
- Manhattan Distance

$$\Delta(V_1, V_2) = \frac{1}{n}\sum_{i=1}^{n} |V_1^i - V_2^i|$$

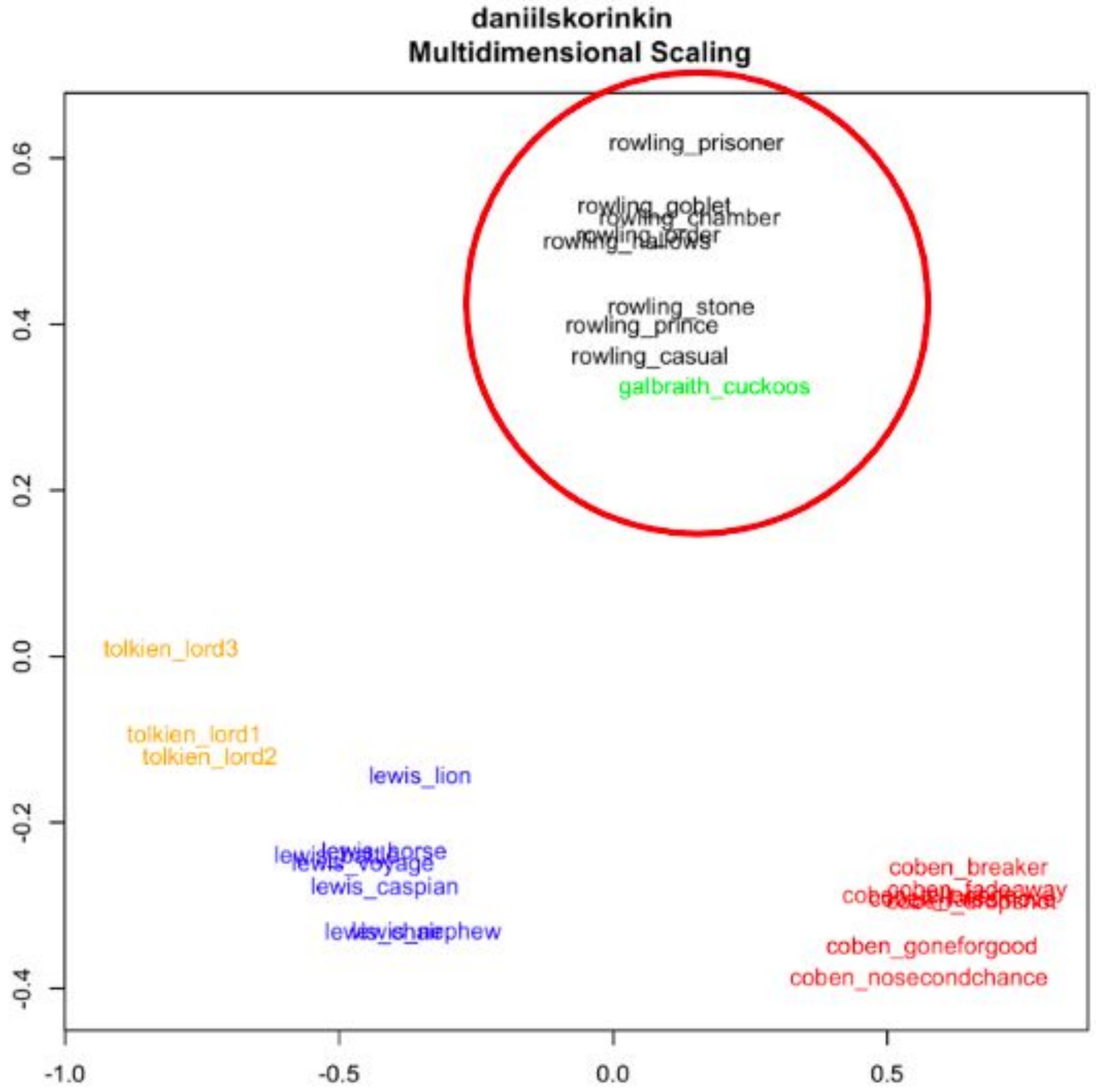


Burrows’ Delta (Burrows 2002)

- Basically: **Manhattan Distance** between two **vectors** of **z-scores**.

$$\Delta(D_1, D_2) = \frac{1}{n}\sum_{i=1}^{n} |z_i(D_1) - z_i(D_2)|$$

Robert Galbraith



Burrows’ Delta

Advantage

- Seems to be working very well with authorship attribution.
- Works across languages.
- See:
  - (Burrows 2002)
  - (Hoover 2004)
  - (Eder, Rybicki 2012)

Disadvantage

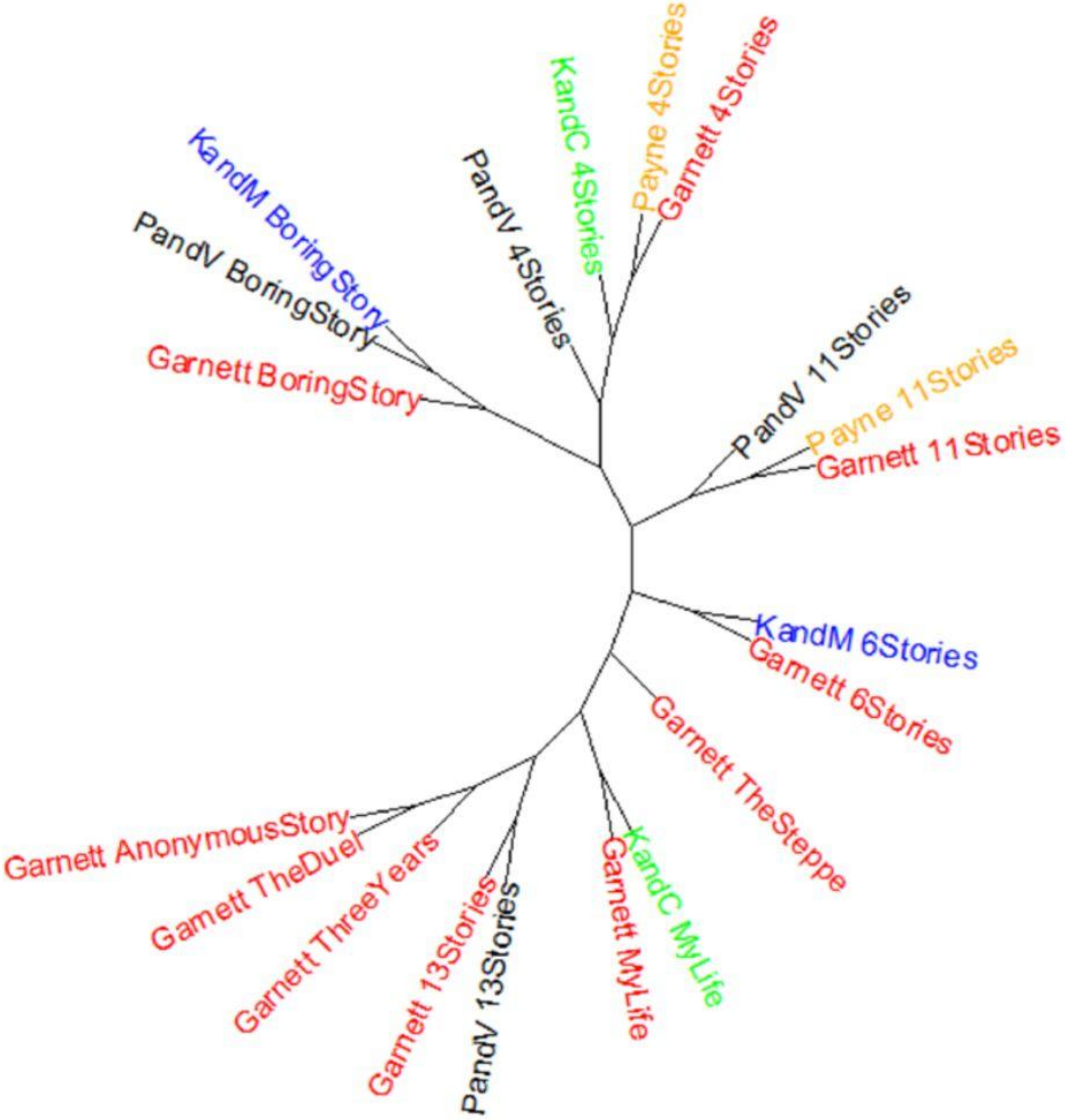
- Not clear **why** it works.

Let’s talk about translation.

Invisible Translator

(Hoover 2019)

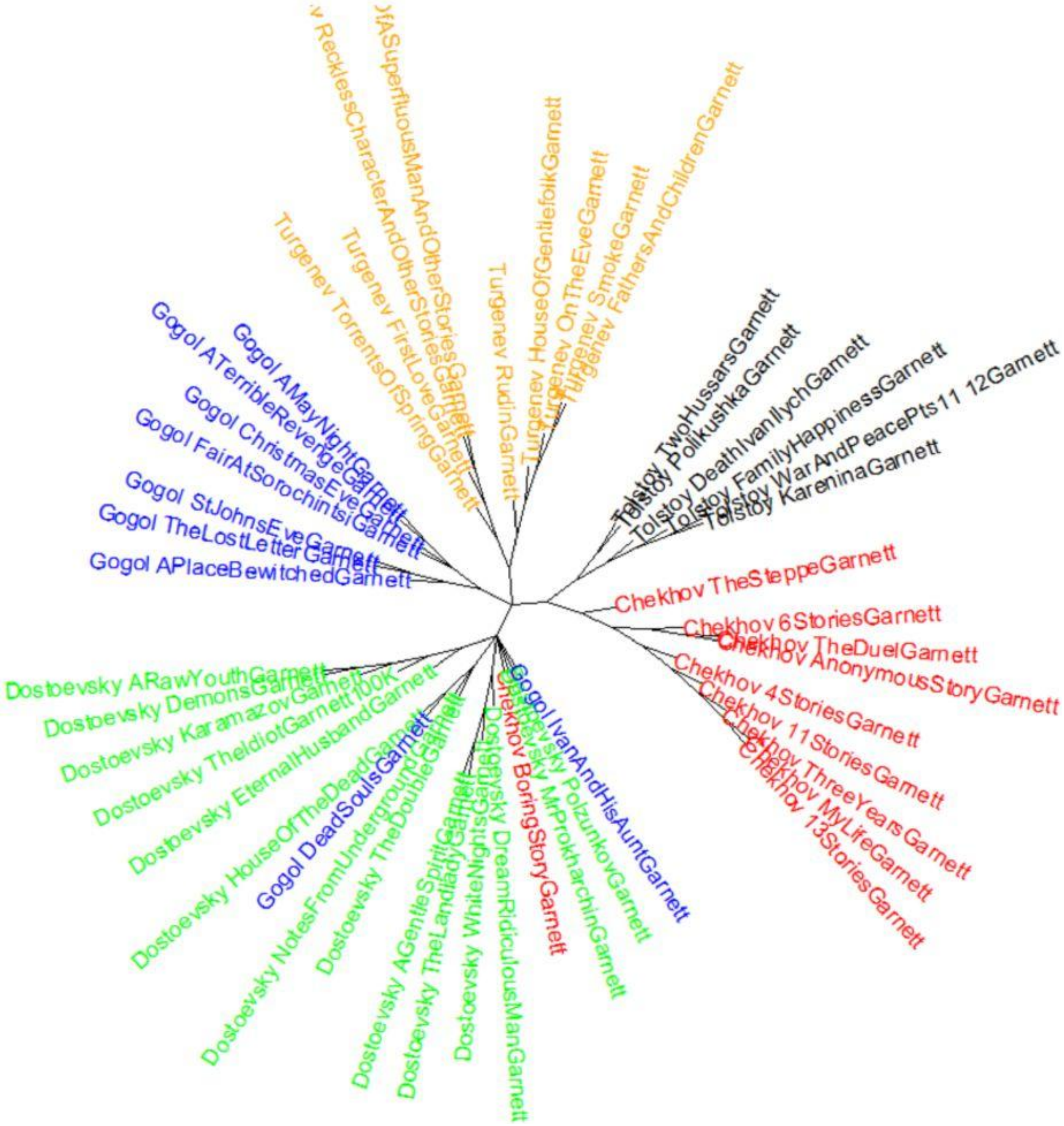
- Is the “signal” of the translator strong?
- Corpus:
  - 1 Russian author
  - 5 English translators
- color = translator
- Strongest signal – text!



Invisible Translator

(Hoover 2019)

- Is the "signal" of the translator strong?
- Corpus:
  - 5 Russian authors
  - 1 English translator
- color = author
- Strongest signal – author!


v RecklessCharacterAndOtherStoriesGarnett
fASuperfluousManAndOtherStoriesGarnett
Turgenev TorrentsOfSpringGarnett
Turgenev FirstLoveGarnett
Turgenev RudinGarnett
Turgenev HouseOfGentlefolkGarnett
Turgenev OnTheEveGarnett
Turgenev SmokeGarnett
Turgenev FathersAndChildrenGarnett
Tolstoy TwoHussarsGarnett
Tolstoy PolikushkaGarnett
Tolstoy DeathIvanIlychGarnett
Tolstoy FamilyHappinessGarnett
Tolstoy WarAndPeacePts11 12Garnett
Tolstoy KareninaGarnett
Chekhov TheSteppeGarnett
Chekhov 6StoriesGarnett
Chekhov TheDuelGarnett
Chekhov AnonymousStoryGarnett
Chekhov 4StoriesGarnett
Chekhov 11StoriesGarnett
Chekhov ThreeYearsGarnett
Chekhov MyLifeGarnett
Chekhov 13StoriesGarnett
Gogol IvanAndHisAuntGarnett
Dostoevsky PolzunkovGarnett
Dostoevsky MrProkharchinGarnett
Chekhov BoringStoryGarnett
Dostoevsky DreamRidiculousManGarnett
Dostoevsky WhiteNightsGarnett
Dostoevsky TheLandladyGarnett
Dostoevsky AGentleSpiritGarnett
Dostoevsky TheDoubleGarnett
Dostoevsky NotesFromUndergroundGarnett
Gogol DeadSoulsGarnett
Dostoevsky HouseOfTheDeadGarnett
Dostoevsky EternalHusbandGarnett
Dostoevsky TheIdiotGarnett
Dostoevsky KaramazovGarnett100K
Dostoevsky DemonsGarnett
Dostoevsky ARawYouthGarnett
Gogol APlaceBewitchedGarnett
Gogol TheLostLetterGarnett
Gogol StJohnsEveGarnett
Gogol FairAtSorochintsiGarnett
Gogol ChristmasEveGarnett
Gogol ATerribleRevengeGarnett
Gogol AMayNightGarnett

Boris Pasternak

- Russian poet. Has translated Shakespeare.
- In his own words:
  - His works "must be judged as original Russian dramatic works" because they have "most of the deliberate freedom without which there is no getting near to great things”
  - The translator has the duty to "to avoid the vocabulary which is not common to them and literary pretentiousness"

Hypothesis

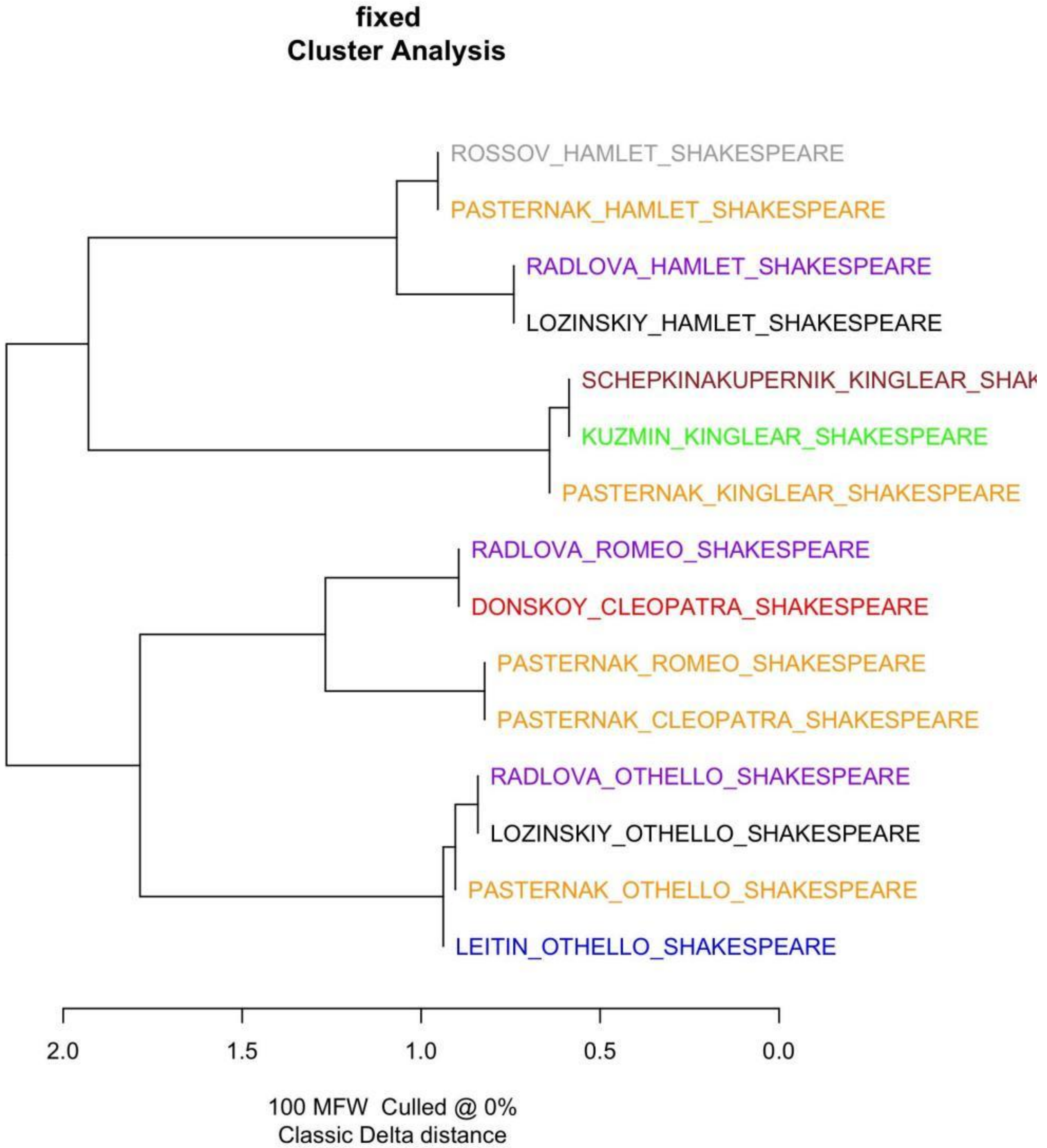
Compared to other Russian translations of Shakespeare,

Pasternak will have a stronger "signal", i.e. Burrows' Delta will be smaller between his translations, compared to other translators of Shakespeare.

Results

At 100 MFW:

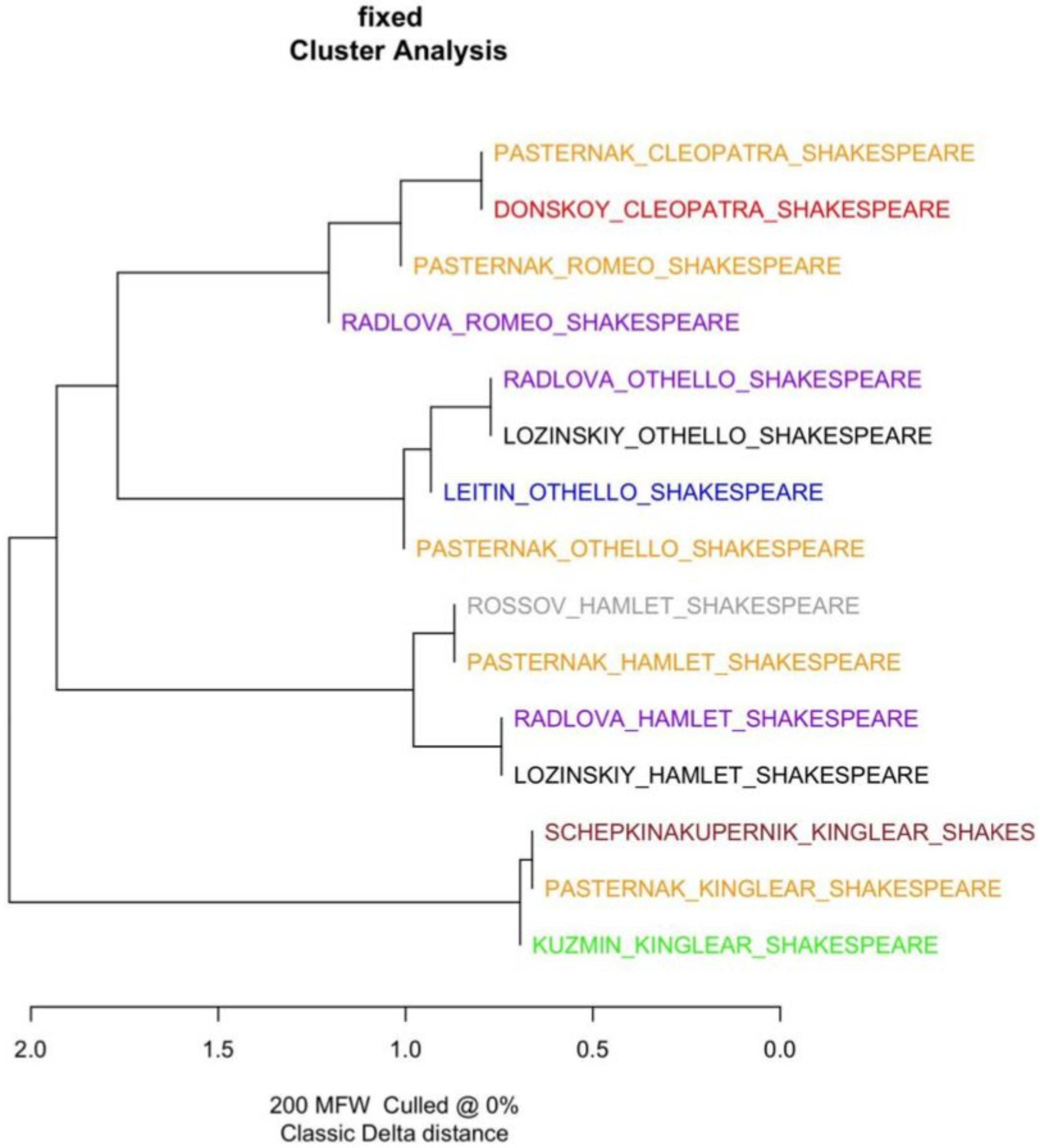
- Pasternak's translations of "Romeo and Juliet" and "Antony and Cleopatra" are closer together than the corresponding translations of Radlova and Donskoy.
- The rest group according to the "text signal" rule.



Results

At 200 MFW:

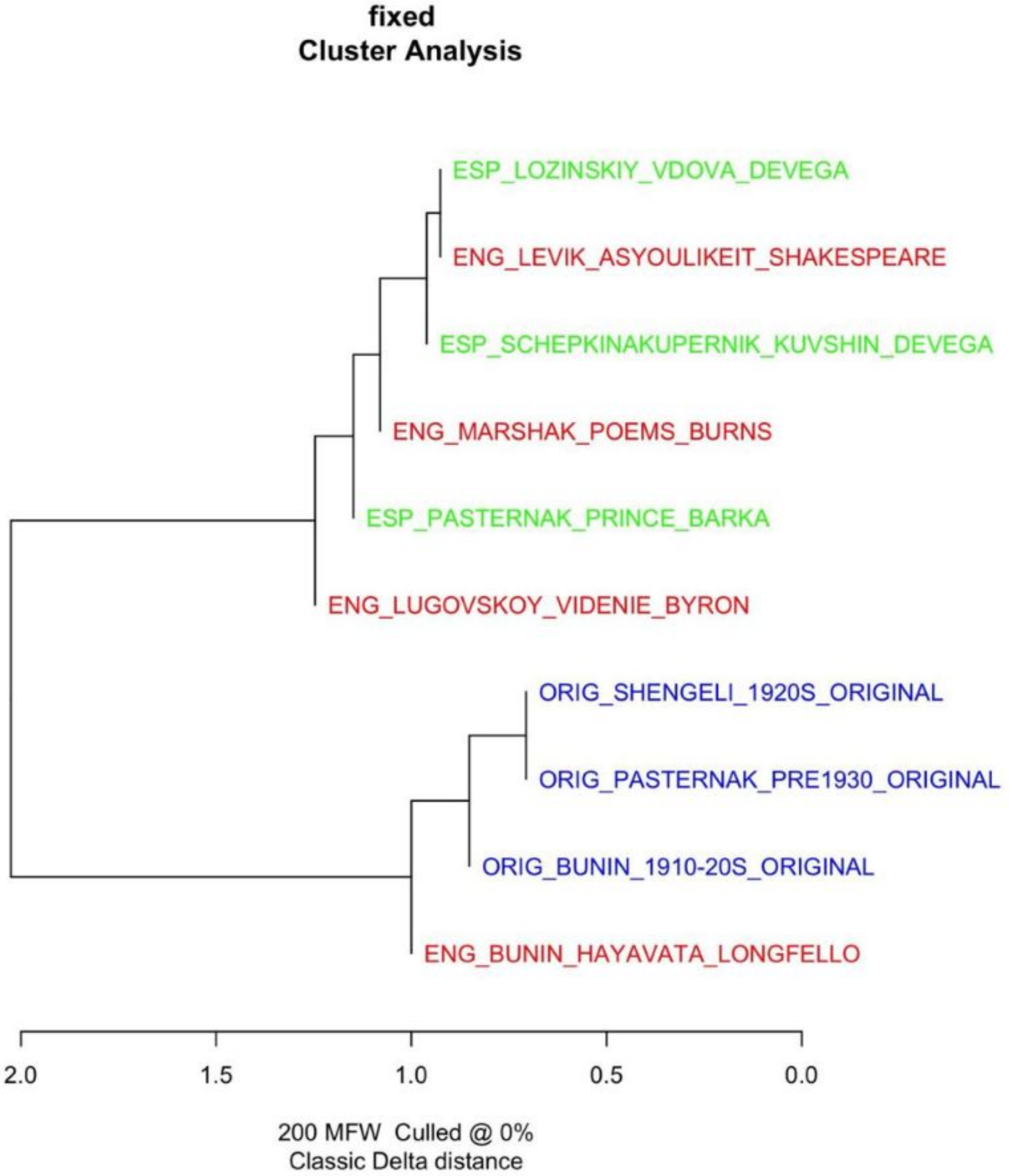
- The magic is gone.
- Pasternak's translations are grouped with other translators' corresponding texts.



Results

Language of the translation doesn't seem to influence the Burrows' Delta.

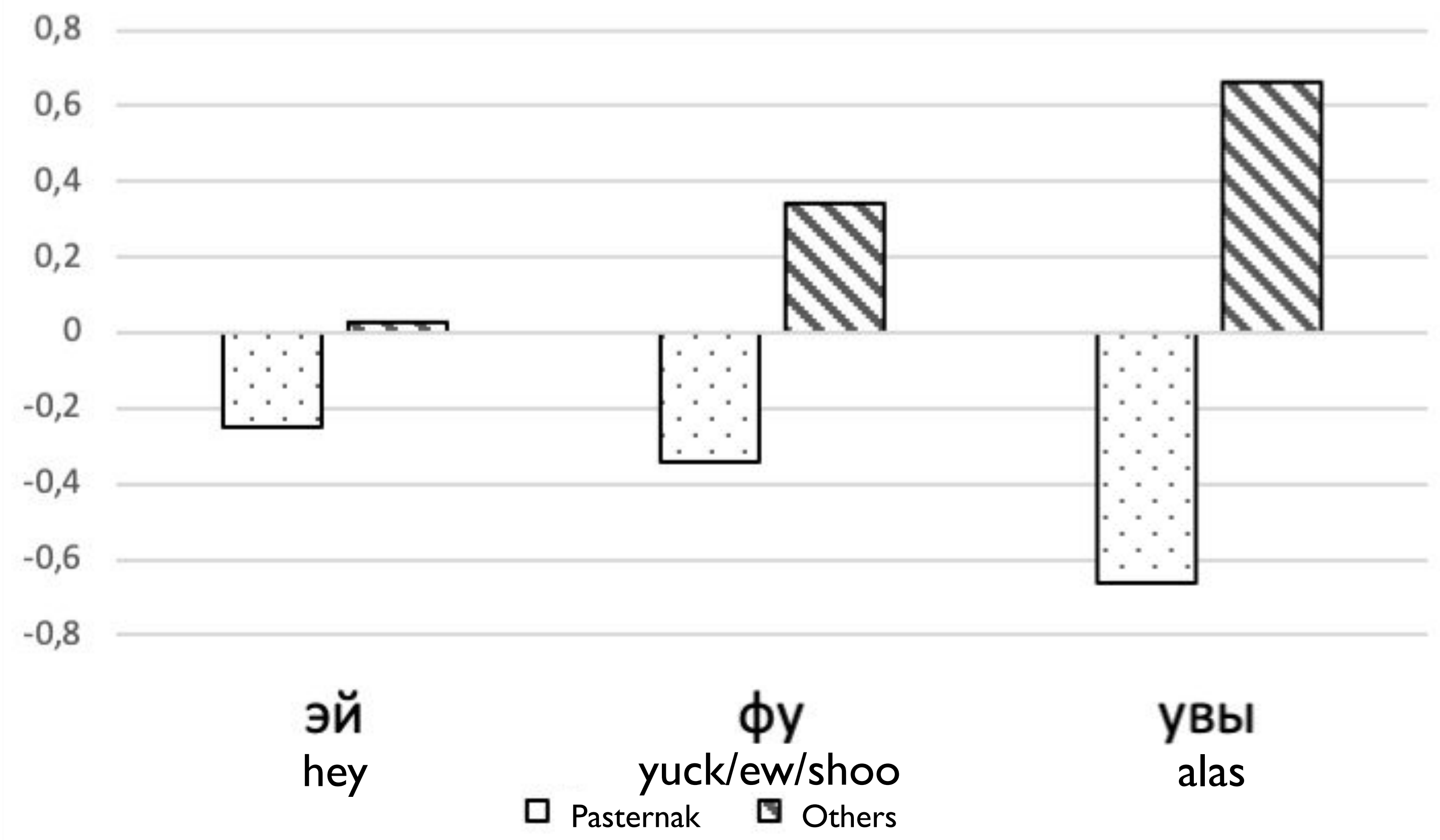
Originally Russian texts are grouped together.



Results

- Out of 100 MFW: three interjections: *эй* (“hey”), *фу* (“ew/yuck/shoo”), *увы* (“alas”)
- Pasternak’s z-scores for all these words are much lower than the z-scores of other translators.
- Does that tell us that Pasternak avoids interjections?

Results: z-scores



Other results

- Pasternak prefers a more common use of "good night".
    - according to the entries in the Russian National Corpus
- Pasternak prefers abbreviated forms of function words.
    - According to some Russian linguists (Dobrushina 2009, Bottineau 2020), some of these forms point to a more "colloquial" manner of speech.
    - See "literary pretentiousness"!

Discussion

- Delta between some of Pasternak's translations is lower at 100 MFW, compared to other translators
  => Pasternak's style is more unique?
- Normalized scores of word frequencies: a valid pattern to look into? What does it tell us about the authors/translators?
- Delta
  - Problematic method, not clear how it works with authorship attribution.
  - Might be very specific to this corpus of translations. (i.e. expensive to verify)

Bibliography

- Bottineau, T. (2020). Opyt sravnitel'nogo opisaniya upotrebleniya uzhe i uzh s pozicii lingvistiki vyskazyvaniya. Russian Linguistics, 44(1).
- Burrows, J. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. Literary and linguistic computing, 17(3), 267-287.
- Dobrushina, N. R. (2009). Semantika častic by i b. Korpusnye issledovanija po russkoj grammatike. M.
- Eder, M., & Rybicki, J. (2012). Do birds of a feather really flock together, or how to choose training samples for authorship attribution. Literary and Linguistic Computing, 28(2), 229-236.
- Hoover, D. L. (2004). Testing Burrows's delta. Literary and linguistic computing, 19(4), 453-475.
- Hoover D. L. (2019). The invisible translator revisited (electronic document). DH2019 official website.
- Liu, H., Chan, R. H., & Yao, Y. (2016). Geometric tight frame based stylometry for art authentication of van Gogh paintings. Applied And Computational Harmonic Analysis, 41(2), 590-602.
- Whissell, C. (1996). Traditional and emotional stylometric analysis of the songs of Beatles Paul McCartney and John Lennon. Computers and the Humanities, 30, 257-265.

Images from

- https://www.npr.org/2017/08/22/545122205/fbi-profiler-says-linguistic-work-was-pivotal-in-capture-of-unabomber
- https://www.labirint.ru/authors/13138/