Count on it,
literature is like a spam

Joanna Byszuk & Jeremi Ochab

DHSI 2024, “DIY Computational Text Analysis with R”

Outline

- Spam: features & filters
- Text classification in literary studies
- Authorship attribution
- Example experiments

How to deal with

SPAM

Dear Friend,

It's my pleasure to Brief you with this proposal for a financial and business assistance. I know my message will come to you as a surprise. Don't worry I was totally convinced to write you in reference to the transfer of $22.5 Million Dollar to your account for onward investment (Hotel industries and Estate building management, Factory and Textile Productions And Extruction of Raw Materials To finished Product For Usage) or any profitable Oriented business in your country.

I Need you to stand as my foreign partner for investment in your country and also next of kin to these fund am about to transfer to you if accepted by you to work with me and receive the fund Amounting to $22.5m.

Please reply immediately if you are interested, so that I can give you more information. Be Rest Assure that these fund transfer to your custody is risk free and profit oriented to both of us.

To enable me start the process and remittance of the fund into your bank account successfully within 10 banking days, I need the following information from you by e-mail: …

May Almighty God Bless You!

Regards, Hanson Chife.

Cześć,

Tutaj jest porównywanie:
https://www.dropbox.com/...

Normalizację robię Gaussem o parametrach wyestymowanych z połączenia obu.

Są podobieństwa, oraz istotne różnice - ciężko coś powiązać z otwartymi oczami, strasznie to skomplikowane.

Chyba najprościej wziąć jakąś sytuację gdzie ICA działa i spróbować poprawić tym.

Pozdrawiam,
Zbychu

Ps. Z rozmiarami wykresów sobie poradziłem dodając "Interpolation".
--
Institute of Computer Science and Computer Mathematics, Jagiellonian University, Cracow, Poland

Dear Friend,

It's my pleasure to Brief you with this proposal for a financial and business assistance. I know my message will come to you as a surprise. Don't worry I was totally convinced to write you in reference to the transfer of **$22.5 Million Dollar** to your account for onward investment (Hotel industries and Estate building management, Factory and Textile Productions And Extruction of Raw Materials To finished Product For Usage) or any profitable Oriented business in your country.

I Need you to stand as my foreign partner for investment in your country and also next of kin to these fund am about to transfer to you if accepted by you to work with me and receive the fund Amounting to $22.5m.

Please reply immediately if you are interested, so that I can give you more information. Be Rest Assure that these fund transfer to your custody is risk free and profit oriented to both of us.

To enable me start the process and remittance of the fund into your bank account successfully within 10 banking days, I need the following information from you by e-mail: …

May **Almighty God Bless You**!

Regards, Hanson Chife.

Cześć,

Tutaj jest porównywanie:
https://www.dropbox.com/...

Normalizację robię Gaussem o parametrach wyestymowanych z połączenia obu.

Są podobieństwa, oraz istotne różnice - ciężko coś powiązać z otwartymi oczami, strasznie to skomplikowane.

Chyba najprościej wziąć jakąś sytuację gdzie **ICA** działa i spróbować poprawić tym.

Pozdrawiam,
Zbychu

Ps. Z rozmiarami wykresów sobie poradziłem dodając "**Interpolation**".
--
Institute of Computer Science and Computer Mathematics, Jagiellonian University, Cracow, Poland

Features of spam

- Look at the e-mail address:
  - known/unknown

Features of spam

- Look at the e-mail address:
  - known/unknown
  - *@uj.edu.pl

Features of spam

- Look at the e-mail address:
  - known/unknown
  - *@uj.edu.pl
- Look for attachments

Features of spam

- Look at the e-mail address:
  - known/unknown
  - *@uj.edu.pl
- Look for attachments
- Look into content:
  - pragmatic: someone wants you to do something
  - semantic: it’s about money
  - register, style, formulaic expressions, etc.

Features of spam

- Look at the e-mail address:
  - ~~known/unknown~~
  - ~~*@uj.edu.pl~~
- Look for attachments
- Look into content:
  - ~~pragmatic: someone wants you to do something~~
  - ~~semantic: it's about money~~
  - ~~register, style, formulaic expressions, etc.~~
  - word occurrences

Spam filter

- detecting spam is a ***classification*** problem
- spam filter is a ***classifier***

Spam filter

- detecting spam is a classification problem
- spam filter is a classifier:
  - collect labelled data
    (spam vs legitimate e-mails)
  - train the model
    (learn   for each class)
  - test it
    (compute   for a message
    and decide)

Spam filter

- detecting spam is a classification problem
- spam filter is a classifier:
  - collect labelled data
    (spam vs legitimate e-mails)
  - train the model
    (learn word probabilities for each class)
  - test it
    (compute word occurrences for a message
    and decide)

Naive **Bayes** spam filter

$$\Pr(S|W) = \frac{\Pr(W|S)\Pr(S)}{\Pr(W|S)\Pr(S)+\Pr(W|H)\Pr(H)}$$

- Pr(S|W) – probability that the message S is a spam, given that it contains word W

M. Sahami, S. Dumais, D. Heckerman, E. Horvitz (1998). "A Bayesian approach to filtering junk e-mail" (PDF). AAAI'98 Workshop on Learning for Text Categorization.

Naive Bayes spam filter

$$\Pr(S|W) = \frac{\Pr(W|S)\Pr(S)}{\Pr(W|S)\Pr(S)+\Pr(W|H)\Pr(H)}$$

- Pr(S|W) – probability that the message S is a spam, given that it contains word W

- 1/5 e-mails are legitimate, the rest is spam:

$$\Pr(S) = 80\%,\ \Pr(H) = 20\%$$

M. Sahami, S. Dumais, D. Heckerman, E. Horvitz (1998). "A Bayesian approach to filtering junk e-mail" (PDF). AAAI'98 Workshop on Learning for Text Categorization.

Naive Bayes spam filter

$$\Pr(S|W) = \frac{\Pr(W|S)\Pr(S)}{\Pr(W|S)\Pr(S)+\Pr(W|H)\Pr(H)}$$

- $\Pr(S|W)$ – probability that the message S is a spam, given that it contains word W

- 1/5 e-mails are legitimate, the rest is spam:

$$\Pr(S) = 80\%, \Pr(H) = 20\%$$

- we need to train $\Pr(W|S)$ and $\Pr(W|H)$ based on known data

M. Sahami, S. Dumais, D. Heckerman, E. Horvitz (1998). "A Bayesian approach to filtering junk e-mail" (PDF). AAAI'98 Workshop on Learning for Text Categorization.

Naive Bayes spam filter

$$\Pr(S|W) = \frac{\Pr(W|S)\Pr(S)}{\Pr(W|S)\Pr(S)+\Pr(W|H)\Pr(H)}$$

- Pr(S|W) – probability that the message S is a spam, given that it contains word W

- 1/5 e-mails are legitimate, the rest is spam:

  $$\Pr(S) = 80\%, \Pr(H) = 20\%$$

- we need to train Pr(W|S) and Pr(W|H) based on known data
- simple threshold criterion, e.g.:
  if $\Pr(S|W_1, \ldots, W_n)$>95% remove that e-mail

M. Sahami, S. Dumais, D. Heckerman, E. Horvitz (1998). "A Bayesian approach to filtering junk e-mail" (PDF). AAAI'98 Workshop on Learning for Text Categorization.

Naive Bayes spam filter

Classify student/teacher by clothing:

$$\boldsymbol{data} = \{standing/sitting, speaking/silent\}$$

$$p(\boldsymbol{teacher}|\boldsymbol{data}) \sim p(\boldsymbol{teacher})p(\boldsymbol{standing}|\boldsymbol{teacher})p(\boldsymbol{speaking}|\boldsymbol{teacher})$$
$$p(\boldsymbol{student}|\boldsymbol{data}) \sim p(\boldsymbol{student})p(\boldsymbol{standing}|\boldsymbol{student})p(\boldsymbol{speaking}|\boldsymbol{student})$$

Choose the class for which the number is higher!

Spam filter summary

- look at features: tokens
- assumption: bag of words („naive" independence)
- correlate with known categories of e-mails
- classify into these (two) categories

How to deal with

LITERATURE

Counting text

- The idea for quantitatively determining authenticity of Pauline epistles by Augustus de Morgan in 1851
- The classics of quantitative text studies is Wincenty Lutosławski's *The origin and growth of Plato's logic: with an account of Plato's style and of the chronology of his writings* from 1897
- The first use of a computing machine (though non-electronic) in stylometry was a study by Thomas C. Mendenhall (1901) A mechanical solution of a literary problem, *Popular Science Monthly* 60
- Father Roberto Busa computationally working on *Index Thomisticus* with IBM starting from 1949

J Rybicki (2014) Pierwszy rzut oka na stylometryczną mapę literatury polskiej. *Teksty Drugie* 2, pp. 106-128

Classification in literature

AUTHOR

Classification in literature

AUTHOR

- What’s Elena Ferrante’s real identity?

Drawing Elena Ferrante's Profile. Workshop Proceedings, Padova, Sept 7, 2017.

- How could one tell Galbraith was Rowling?

P Juola (2013). How a Computer Program Helped Show J.K. Rowling write A Cuckoo’s Calling. *Scientific American,* Aug 20, 2013.

Classification in literature

AUTHOR

- Authorial collaborations – who’s writing and who’s editing?

J Rybicki, M Kestemont, D Hoover (2013). Collaborative authorship: Conrad, Ford and rolling Delta. *Digital Humanities 2013: Conference Abstracts*. Lincoln: University of Nebraska-Lincoln, 368-71

- Are both *Go set a watchman* and *To kill a mockingbird* Harper Lee’s?

M Eder, J Rybicki (2015). Go Set A Watchman while we Kill the Mockingbird In Cold Blood
https://sites.google.com/site/computationalstylistics/projects/lee_vs_capote
E Gamerman (2015). Data Miners Dig for Answers About Harper Lee, Truman Capote and *Go Set a Watchman*. *Wall Street Journal,* Jul 15, 2015.

Classification in literature

AUTHOR

TRANSLATOR

Classification in literature

AUTHOR **TRANSLATOR**

- Is the translator invisible? Is the authorial fingerprint retained in translation?

J Rybicki (2013). Stylometryczna niewidzialność tłumacza. *Przekładaniec* 27, 61–87.
J Rybicki (2013). The great mystery of the (almost) invisible translator. In: MP Oakes & M Ji (Eds.) *Quantitative Methods in Corpus-Based Translation Studies*.

- How about the translator's spouse?

J Rybicki (2011). Alma Cardell Curtin and Jeremiah Curtin: the translator's wife's stylistic fingerprint. *Digital Humanities 2011: Conference Abstracts*. Stanford University, Stanford, pp. 308-11.

Classification in literature

AUTHOR **TRANSLATOR**

- Or when did a translator die?

J Rybicki and M Heydel (2013). The stylistics and stylometry of collaborative translation: Woolf's 'Night and Day' in Polish. *Literary and Linguistic Computing*, 28(4): 708-17

- How many scribes helped in *Queen Sophia's Bible* translation?

M Eder (2016). Rolling stylometry. *Digital Scholarship in the Humanities*, 31(3): 457-469

Classification in literature

AUTHOR

TRANSLATOR

LANGUAGE

Classification in literature

AUTHOR

GENDER

TOPIC

GENRE

NARRATION TYPE

LITERATURE PERIOD

LITERATURE MOVEMENT

TRANSLATOR

LANGUAGE

Classification in literature

When classes are defined:

- what are the **distinguishing features** (phrases, syntactic structures, themes, emotional cues, plot shapes, mannerisms)?
- do they change in **time**?
- can one interpret what they serve?

Counting text

Digitisation age:

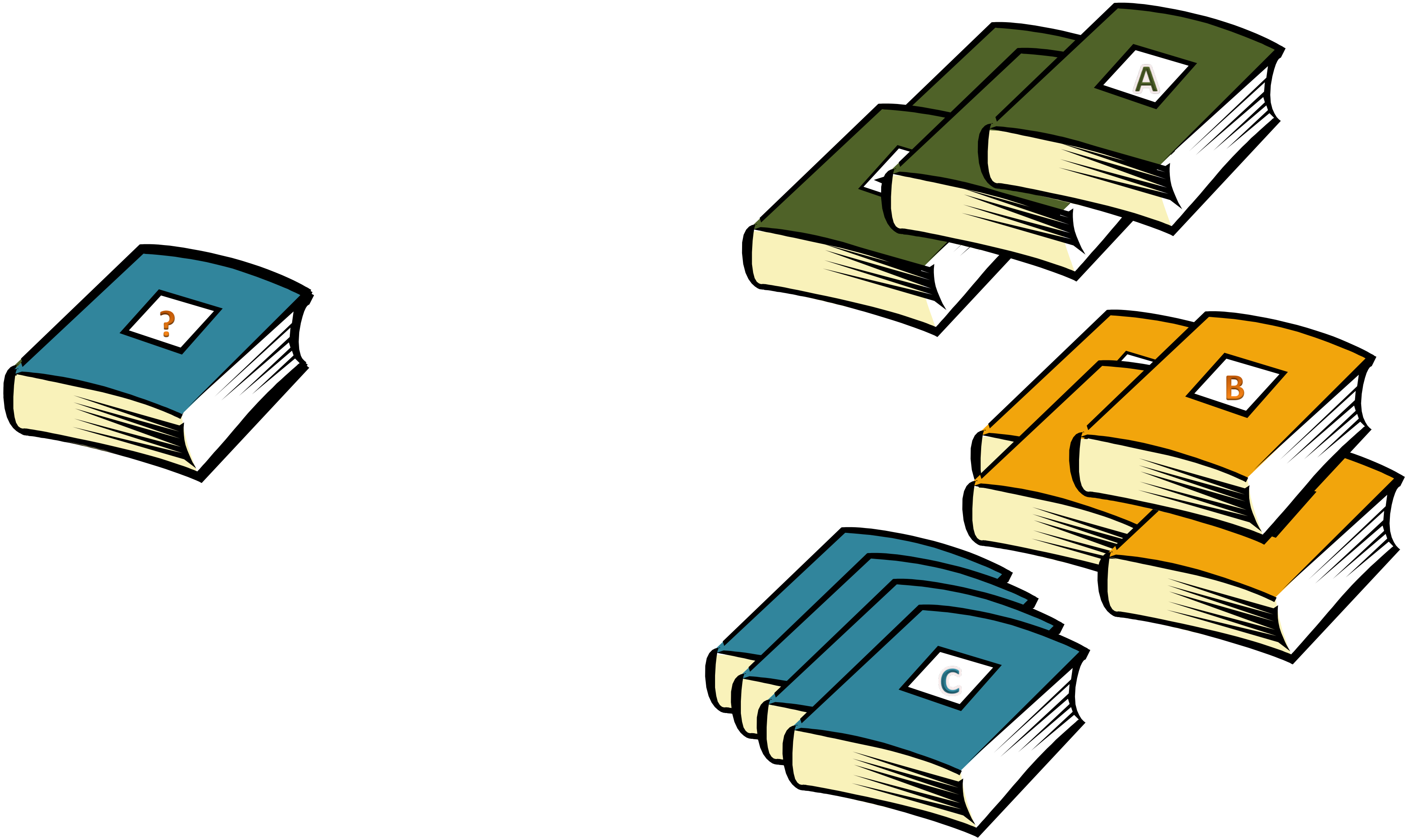
- text recognition software (OCR, HTR)
- digital library archives
- new content is digital by origin

J-B Michel et al. (2011) Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* 331

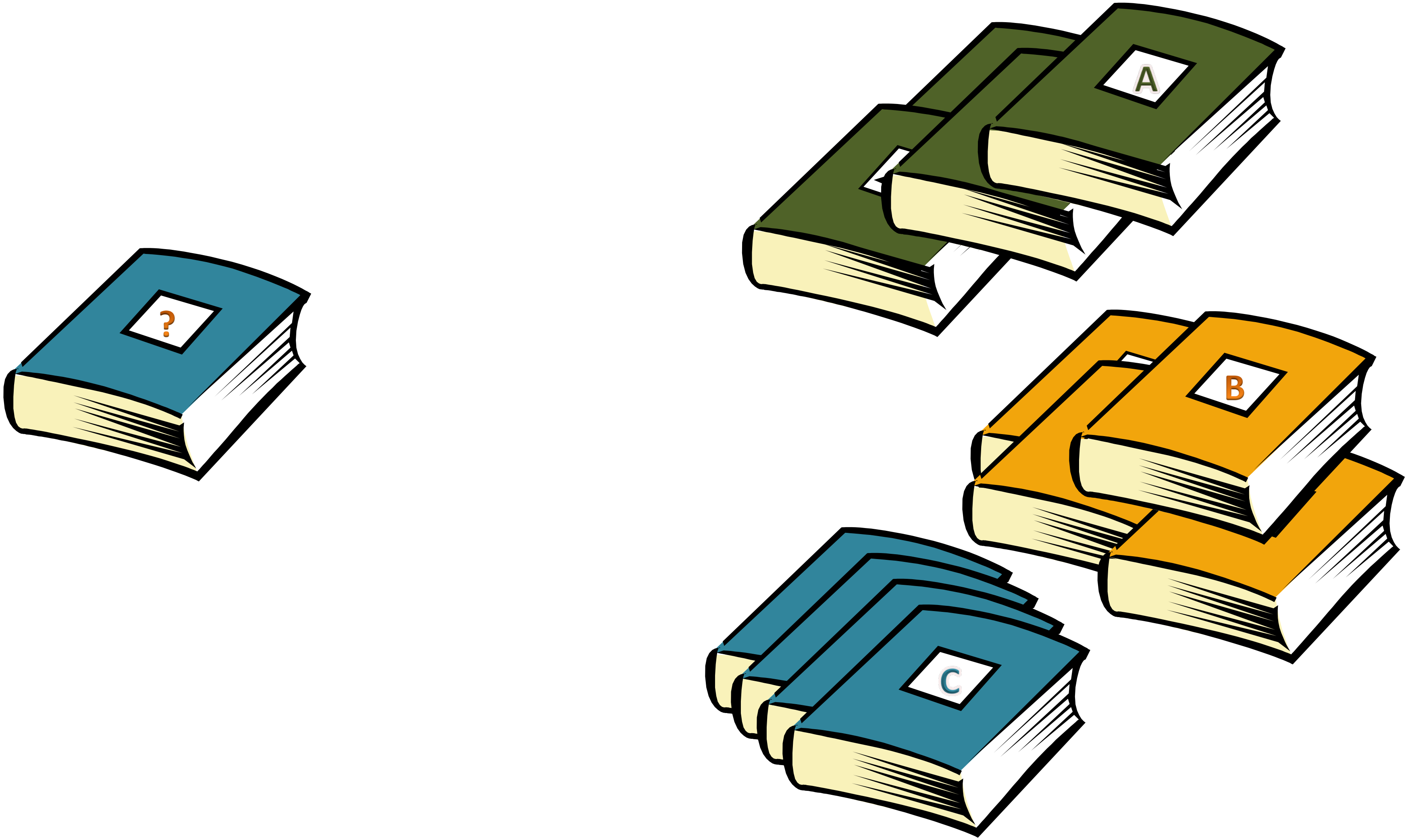
What are the

PROBLEMS WITH TEXT CLASSIFICATION

Example: authorship attribution



Example: authorship attribution



Text closeness

Frequencies of:

- characters
- words
- sentences?
- POS-tags

Other:

- Sentence lengths
- Word lengths

➡

- character N-grams
- word N-grams
- POS-tags N-grams

Text closeness

1. Calculate frequencies of words
2. Calculate distances

words	Book A	Book B
a	120	115
the	100	110
of	70	80
…	…	…

We get one number (distance) for each pair

	Book A	Book B	…
Book A	0	d(A,B)	…
Book B	d(B,A)	0	…
…	…	…	…

Text closeness

But:

- Which words/POS should we take?
- How many of them?
- Should we lemmatise them?

Text closeness

But:

- Which words/POS should we take?
- How many of them?
- Should we lemmatise them?

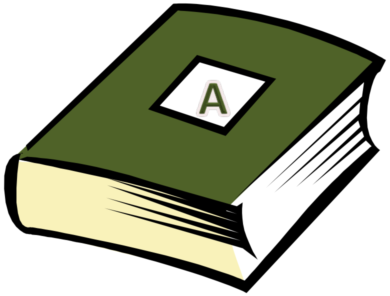
Even before that:

- Which authors should we compare to?
- Which books? How many? How long?

EXPERIMENT 1

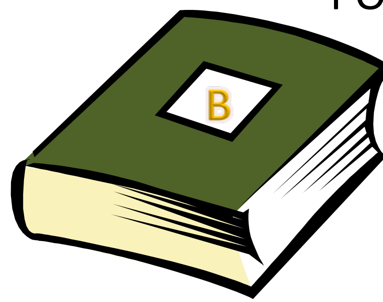
IS GRAMMAR OR VOCABULARY AUTHORIAL?

Fabricate a fake text



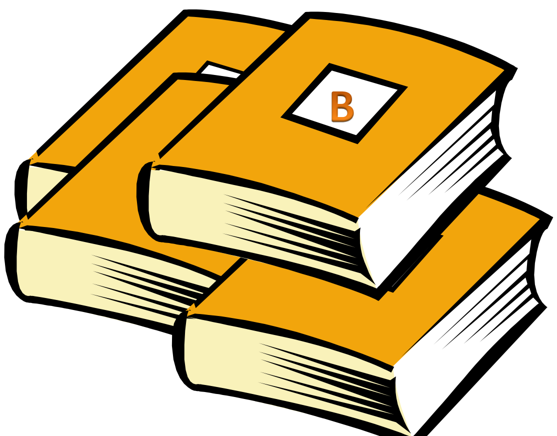
Extract frequencies of POS-tags:
2 IN, 2 DT, 2 NN, 1 VB…

Put words from B in the places of POS-tags from **A.**



of a sate the estate with his time and perpetually , she swim gloomily cultiveated to birthday .

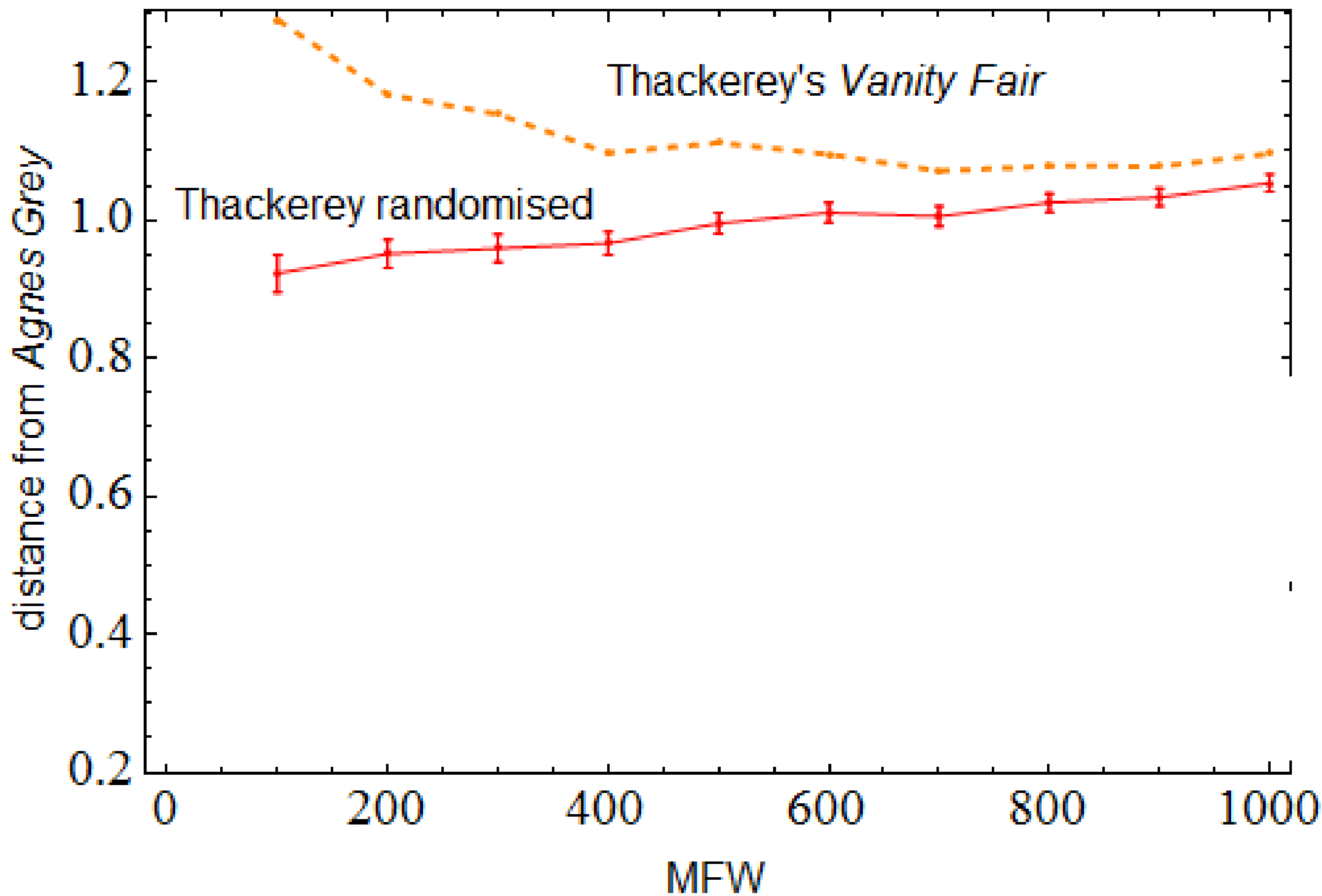
Whether_IN this_DT be_VB the_DT case_NN with_IN my_PRP$ history_NN or_CC not_RB ,_, I_PRP am_VBP hardly_RB competent_JJ to_TO judge_NN ._.



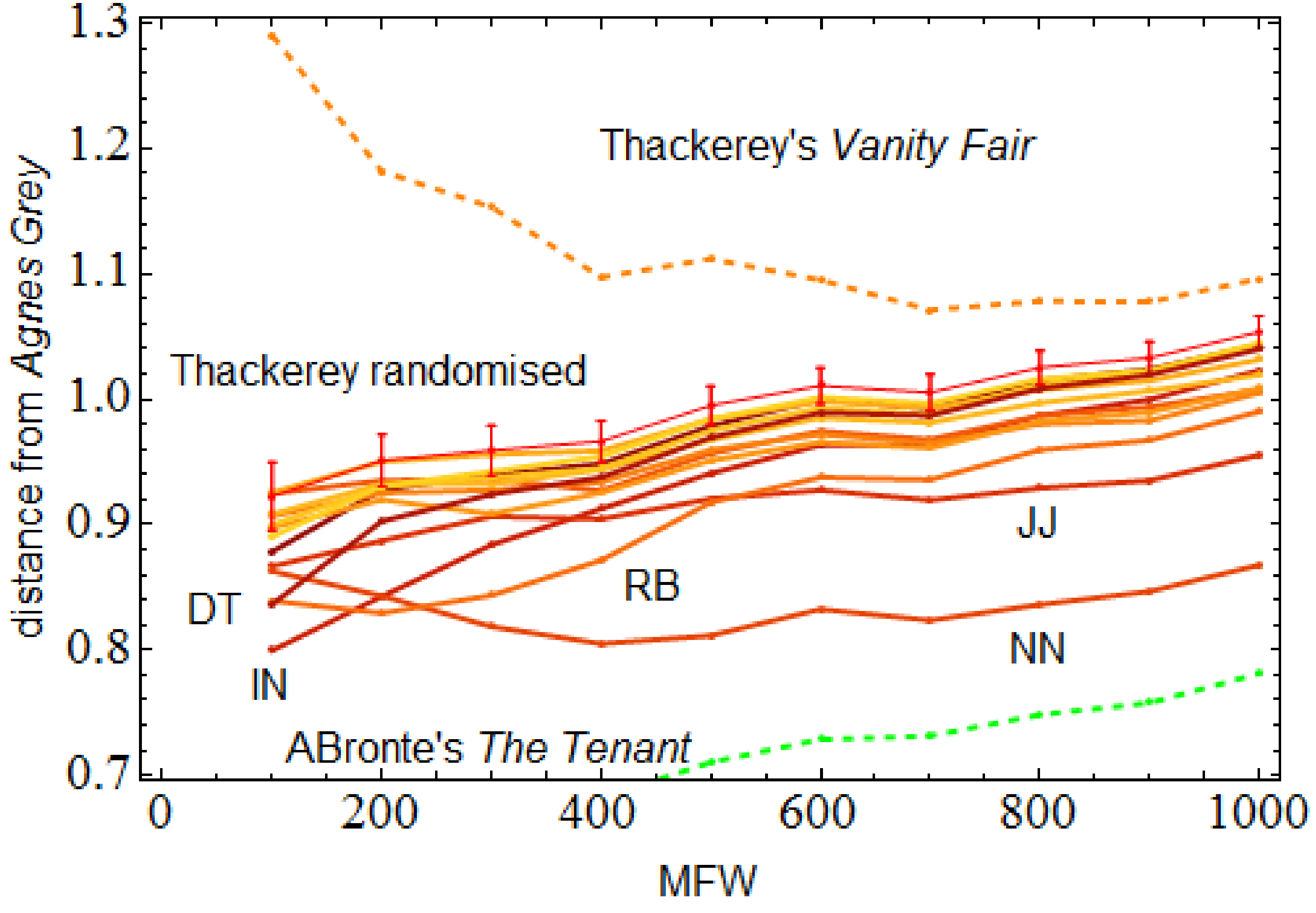
Extract words with given POS-tags:
IN → whether, with, …
DT→ this, the, …
…

Stanford tagger: K. Toutanova, D. Klein, C. Manning, and Y. Singer. 2003. ”Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network”. In *Proceedings of HLT-NAACL 2003*, pp. 252-259.

Fabricate a fake text



Fabricate a fake text



Take-home message:

- stylometric distance depends on both **vocabulary** and (implicitly) **syntax**

EXPERIMENT 2: noise

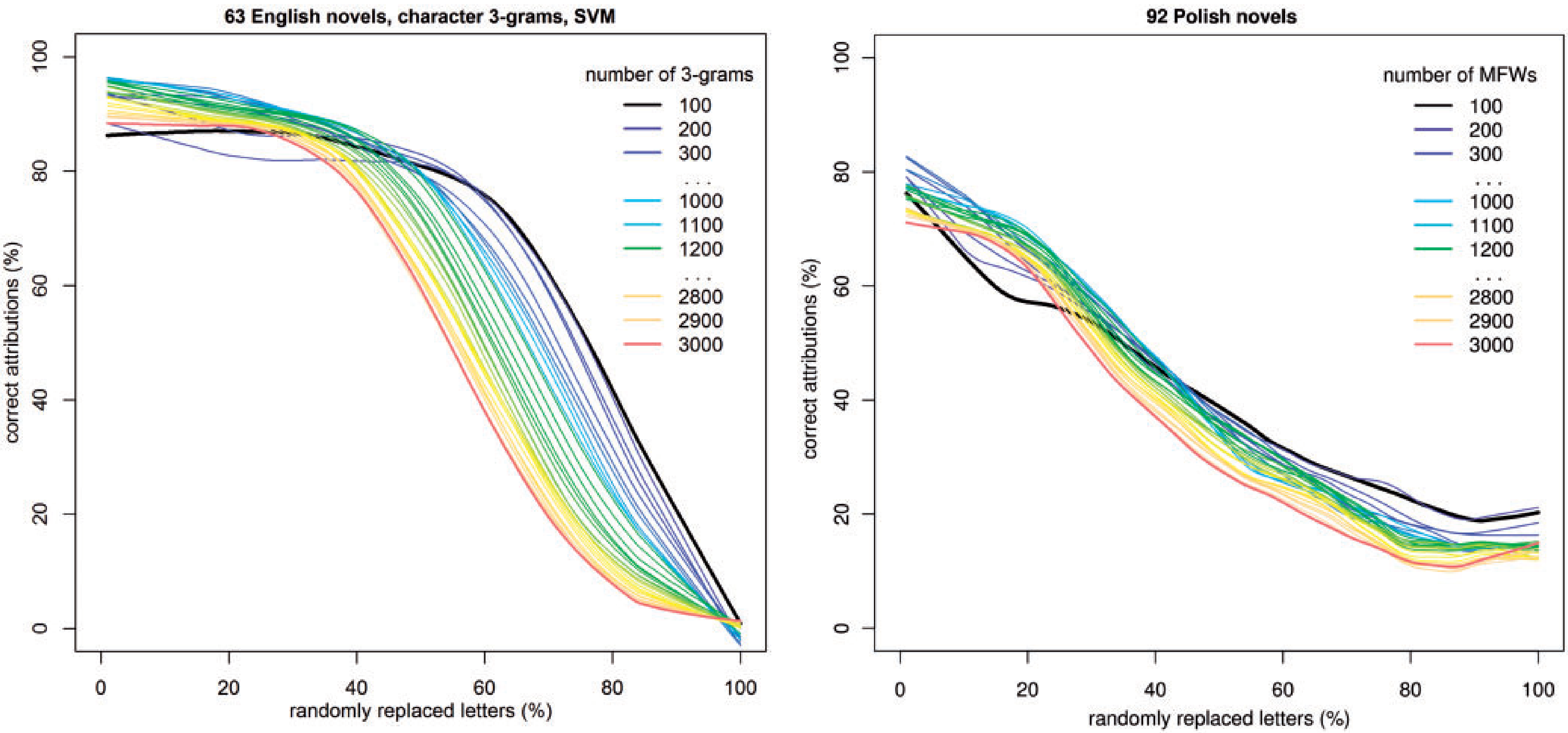
DOES DIGITISATION SPOIL ATTRIBUTION?

Digitisation:

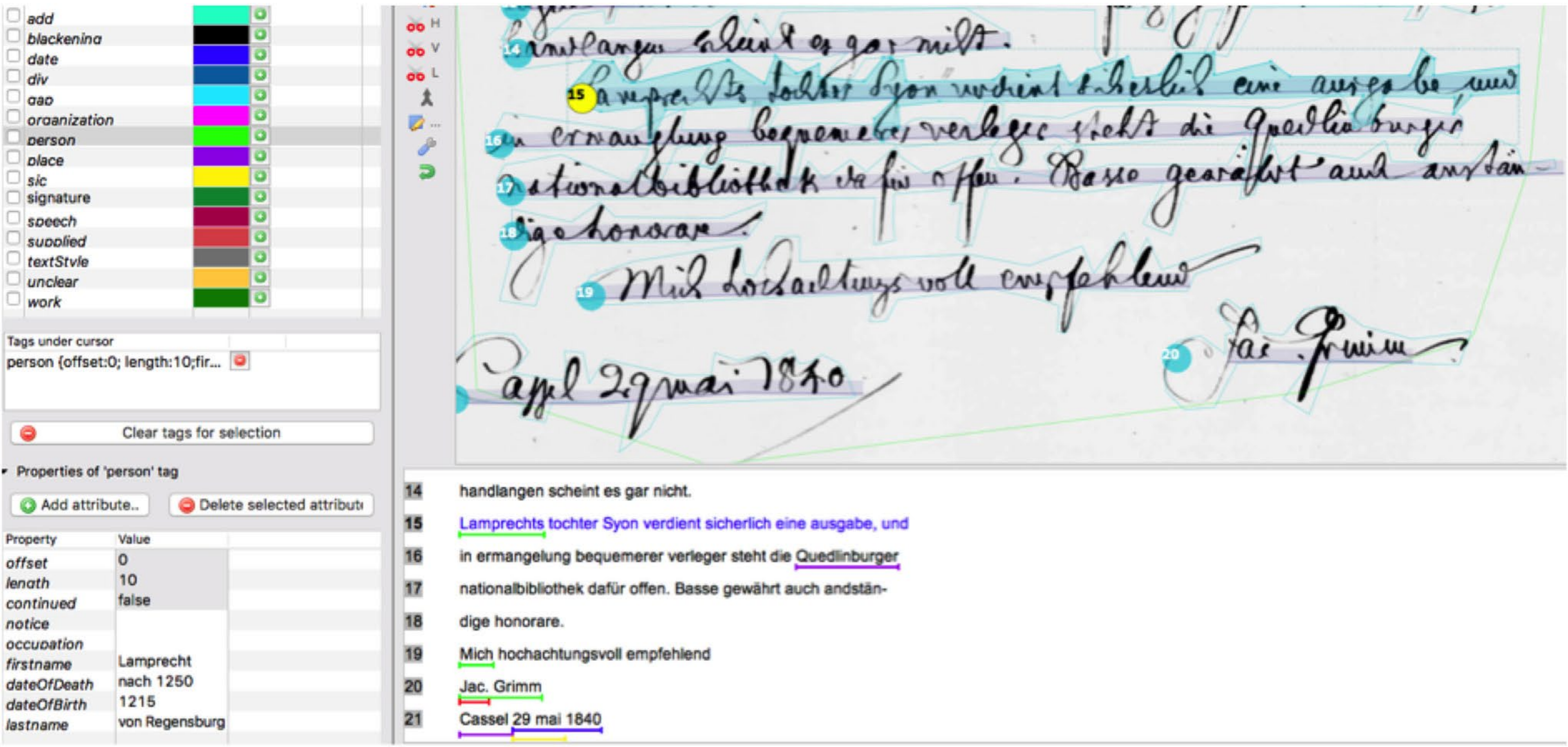
- Optical Character Recognition (OCR)
- Handwritten Text Recognition (HTR)
- Gold standard: human expert transcription

H. Benzerroug, S. Khennouf (2017). Author identification of corrupted OCR-based texts. HDSKD journal 3 (2) pp. 91-99

M. Eder (2013). Mind your corpus: systematic errors in authorship attribution. *Literary and Linguistic Computing*, *28*(4), 603-614.



M. Eder (2013). Mind your corpus: systematic errors in authorship attribution. *Literary and Linguistic Computing*, *28*(4), 603-614.

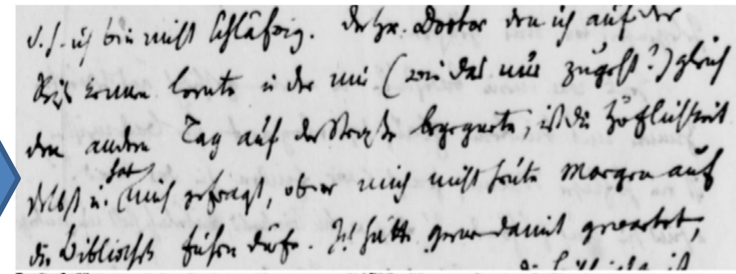


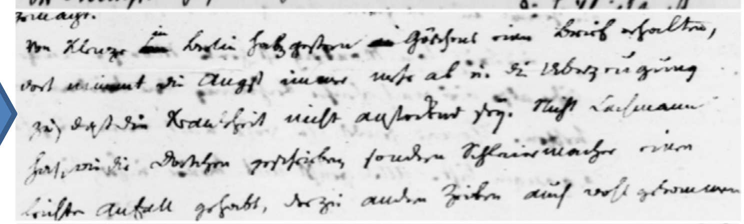
[Image reproduced with permission of the Hessisches Staatsarchiv Marburg].
Jander, M. (2016). *Handwritten Text Recognition – Transkribus: A User Report*. Göttingen, Germany: eTRAP Research Group, University of Göttingen.

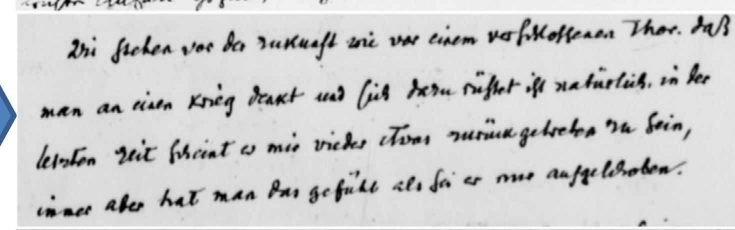
Legibility and cleanliness

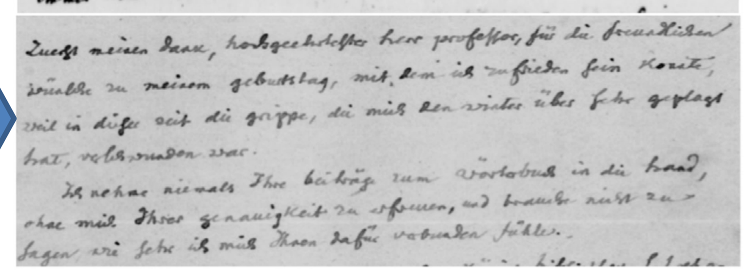
Wilhelm Grimm's letters:

- very low legibility
  (Br 5993, 7 years old)
- low legibility
  (Br 2680, 45 years old)
- medium legibility
  (Br 2743, 73 years old)
- high legibility
  (Br 2736, 69 years old)







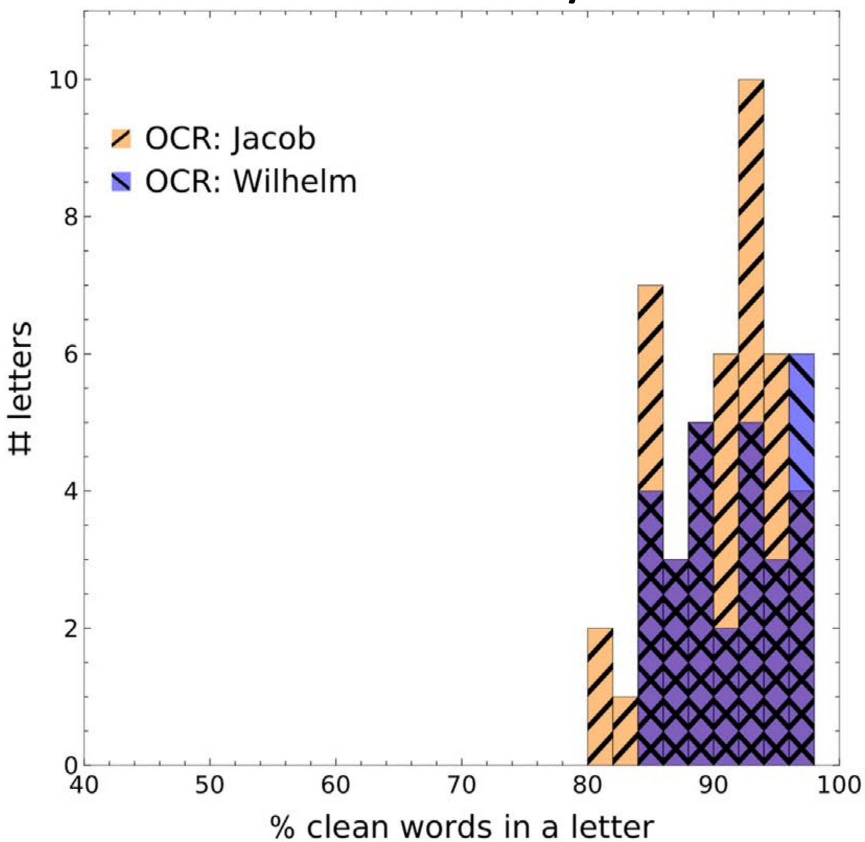


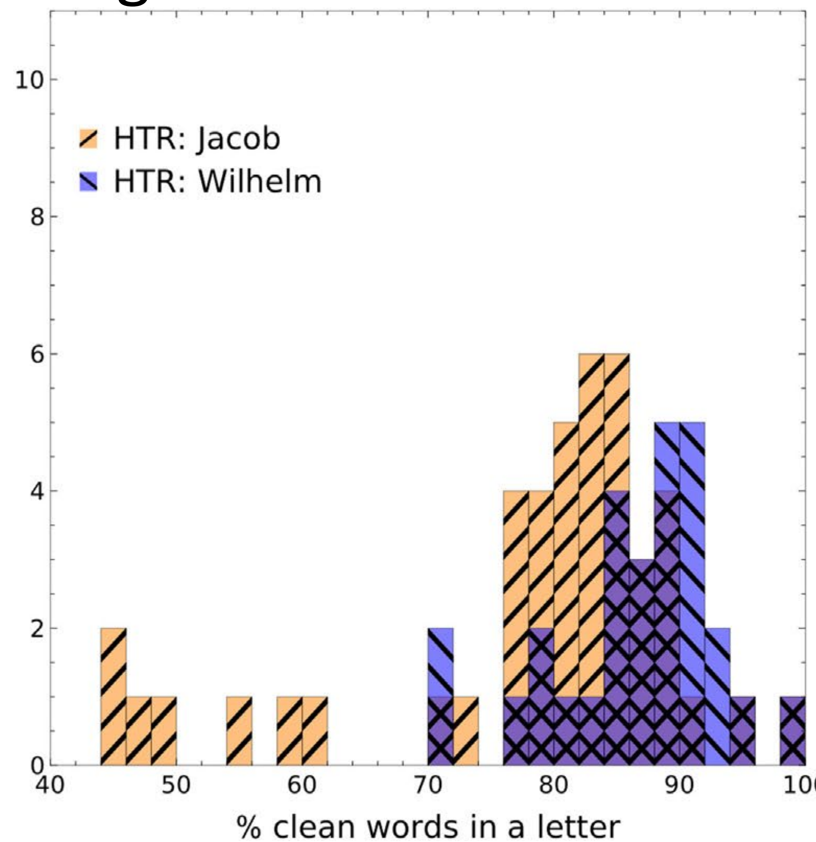
[Images reproduced with permission of the Hessisches Staatsarchiv Marburg].

Legibility and cleanliness

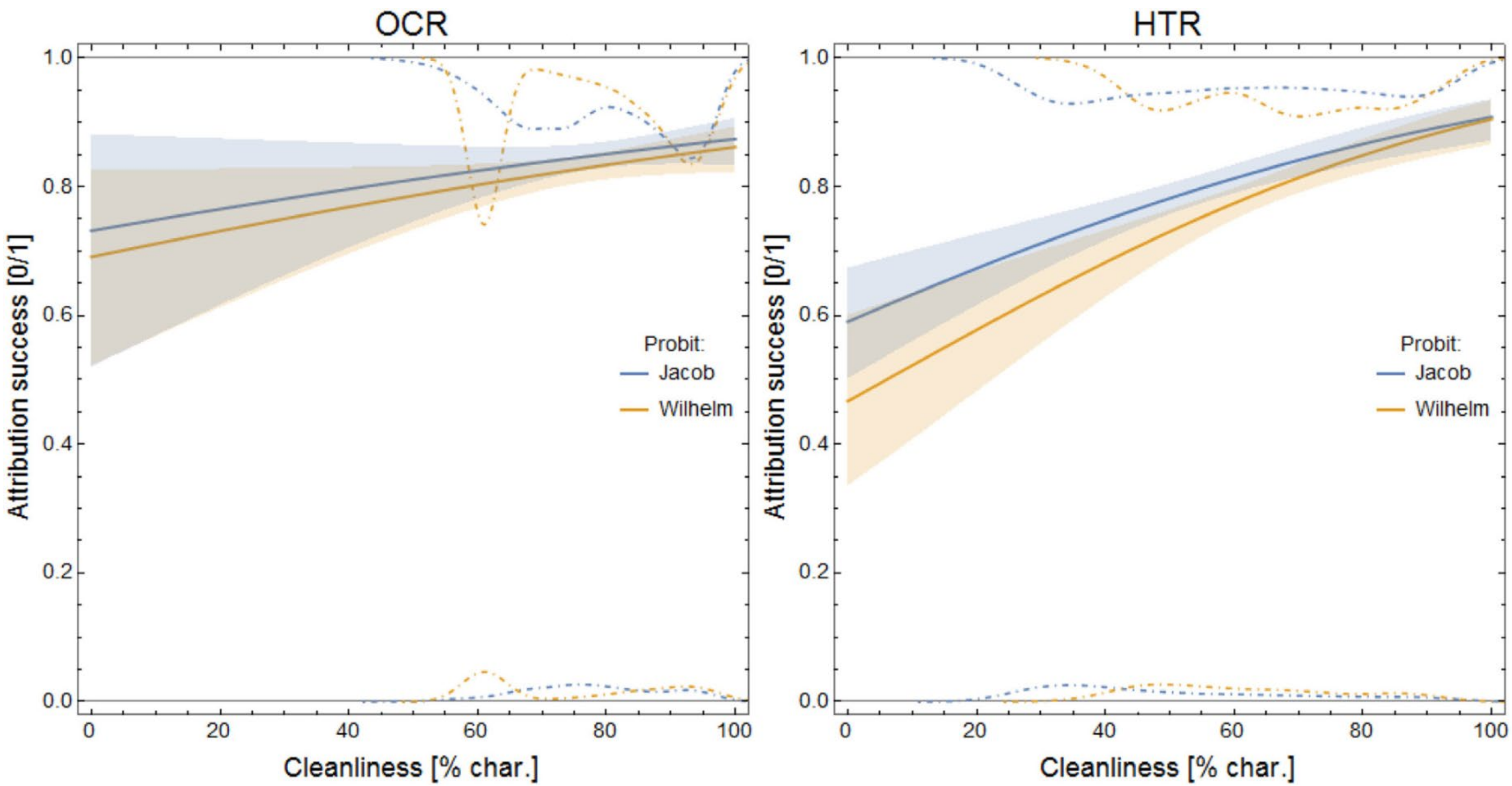
Human-assessed legibility (very low excluded):

- Jacob: 36 low/9 medium–high
- Wilhelm: 15 low/13 medium–high





	MAN	OCR	HTR
Accuracy	91.66	91.66	88.88
F1 score	88.46	88.46	84.61



Franzini G, et al. (2018) Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm. *Front. Digit. Humanit*. **5**:4

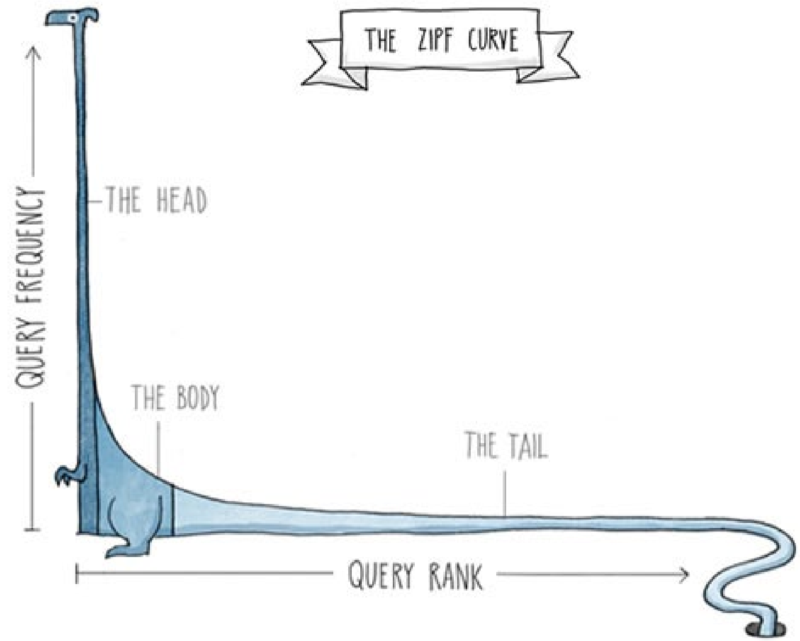
Take-home message:

- significant relation between auth. attr. **performance** and **cleanliness** for **HTR**
- auth. attr. **performs** as **well** on **OCR** as on human transcription

Lexical richness

Stylometry and authorship attribution based on: words/word n-gram/character n-gram frequencies.

How is the **word frequency distribution** affected by errors in HTR/OCR?



Lexical richness

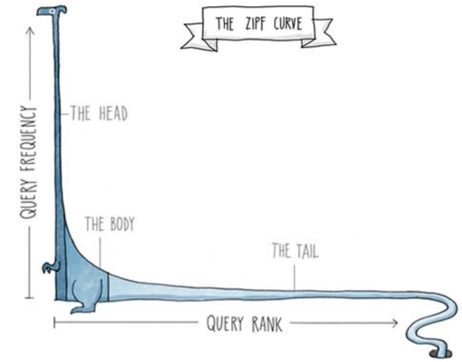
Out of many diversity indices:

- Shannon entropy $H = -\sum_{t=1}^{T} p_t \log p_t$
- Simpson's index (inverse participation ratio) $D = \sum_{t=1}^{T} p_t^2$

Simple, least arbitrary, theoretically understood, known limiting values

Tweedie, F.J., and Baayen, R.H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities* 32: 323–52. doi:10.1023/A:1001749303137

Lexical richness



Out of many diversity indices:

- Shannon entropy (tails) $H = -\sum_{t=1}^{T} p_t \log p_t$
- Simpson's index (core) $D = \sum_{t=1}^{T} p_t^2$
  (inverse participation ratio)

Simple, least arbitrary, theoretically understood, known limiting values

Tweedie, F.J., and Baayen, R.H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities* 32: 323–52. doi:10.1023/A:1001749303137

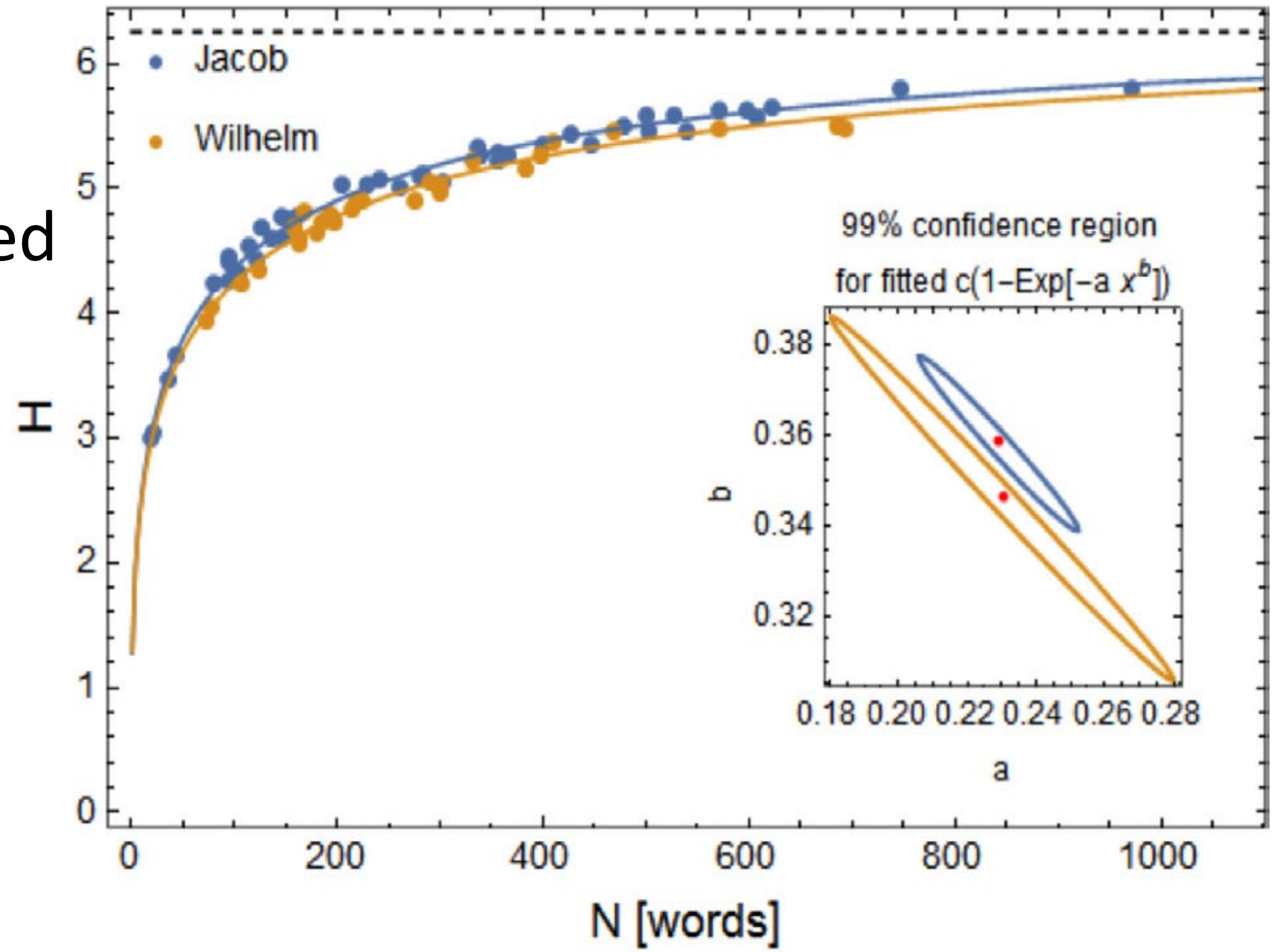
Lexical richness

- HTR produces enough errors to significantly yield lower richness per letter (word omission or merging)
- in short letters probably caused by HTR omitting or merging words
- no other correlations between text richness and cleanliness of HTR or OCR

OCR is more viable for stylometric measurements.

Lexical richness

- can be authorial marker, but…
- depends on **text length**
- can be modelled



EXPERIMENT 3: the temporal

ONE STEP BEYOND BAG OF WORDS

Until now only:

- character
- word
- POS-tags

and their N-grams

Until now only:

- character
- word
- POS-tags

and their N-grams

But text is comprised of symbolic **sequences**.

S. Drożdż, et al., arXiv:1412.8319 [cs.CL] (2014).
E. G. Altmann, G. Cristadoro, and M. D. Esposti, PNAS 109, 11582 (2012).

Until now only:

- character
- word
- POS-tags

and their N-grams

But text is comprised of symbolic **sequences**.

Imagine DNA or heart rate time series.

C. Peng, et al., Nature 356, 168 (1992).

Until now only:

- character
- word
- POS-tags

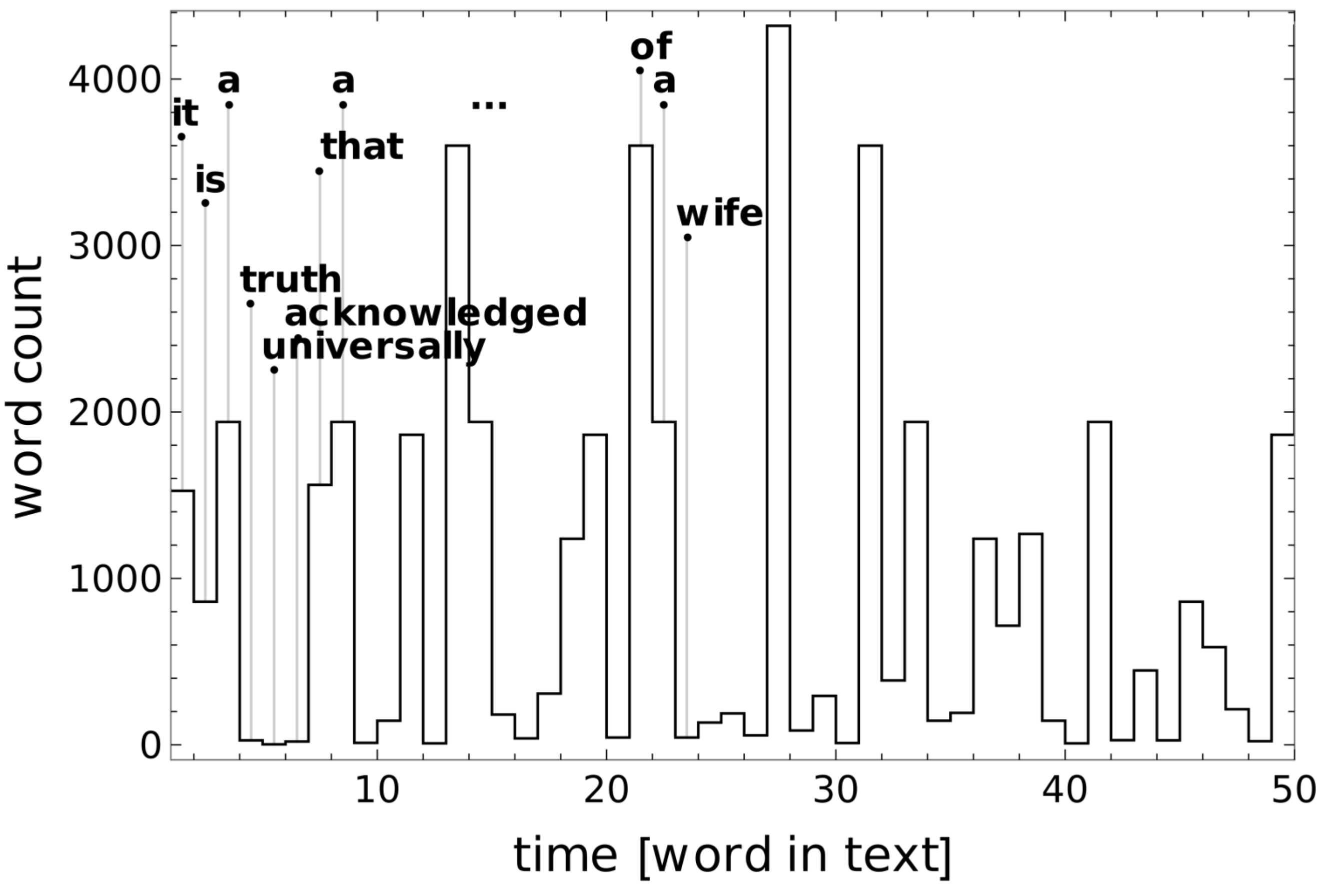
and their N-grams

But text is comprised of symbolic **sequences**.

take texts » ~~count words~~ » learn machines

quantify **change**

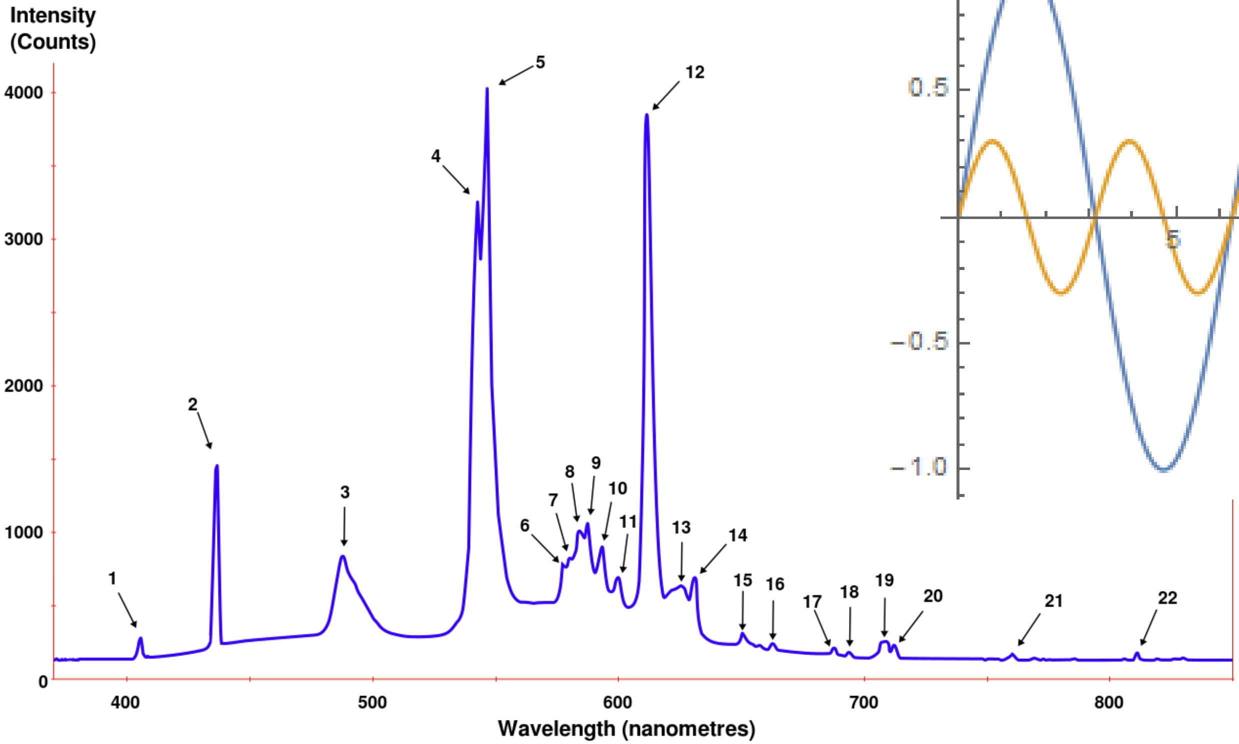
Sequence of ranks

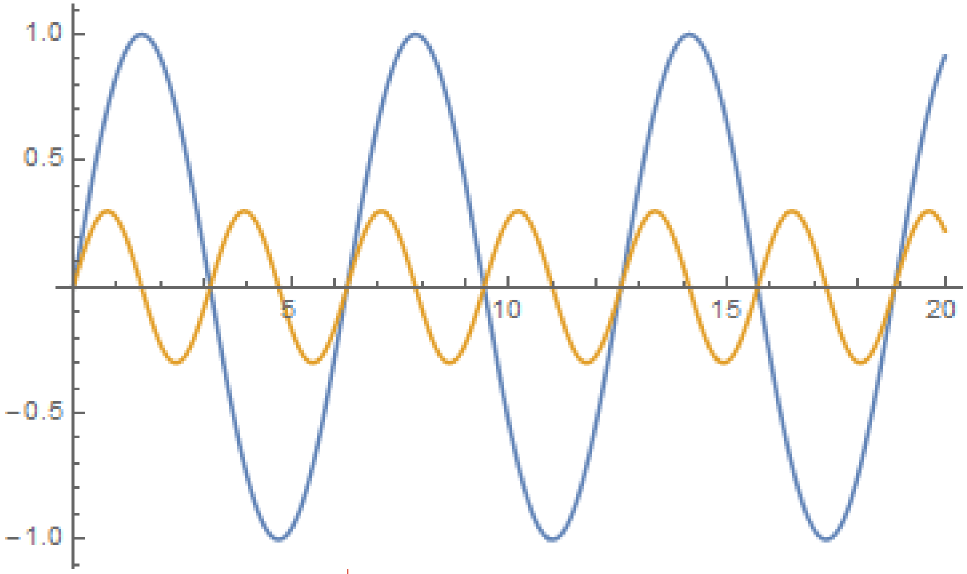


M. A. Montemurro and P. A. Pury, Fractals 10, 451 (2002).
M. Ausloos, Phys. Rev. E 86, 031108 (2012).

Power spectrum of time series

$$S(f) = |\sum_t x_t \exp(-2\pi i f t)|^2 \qquad (1)$$

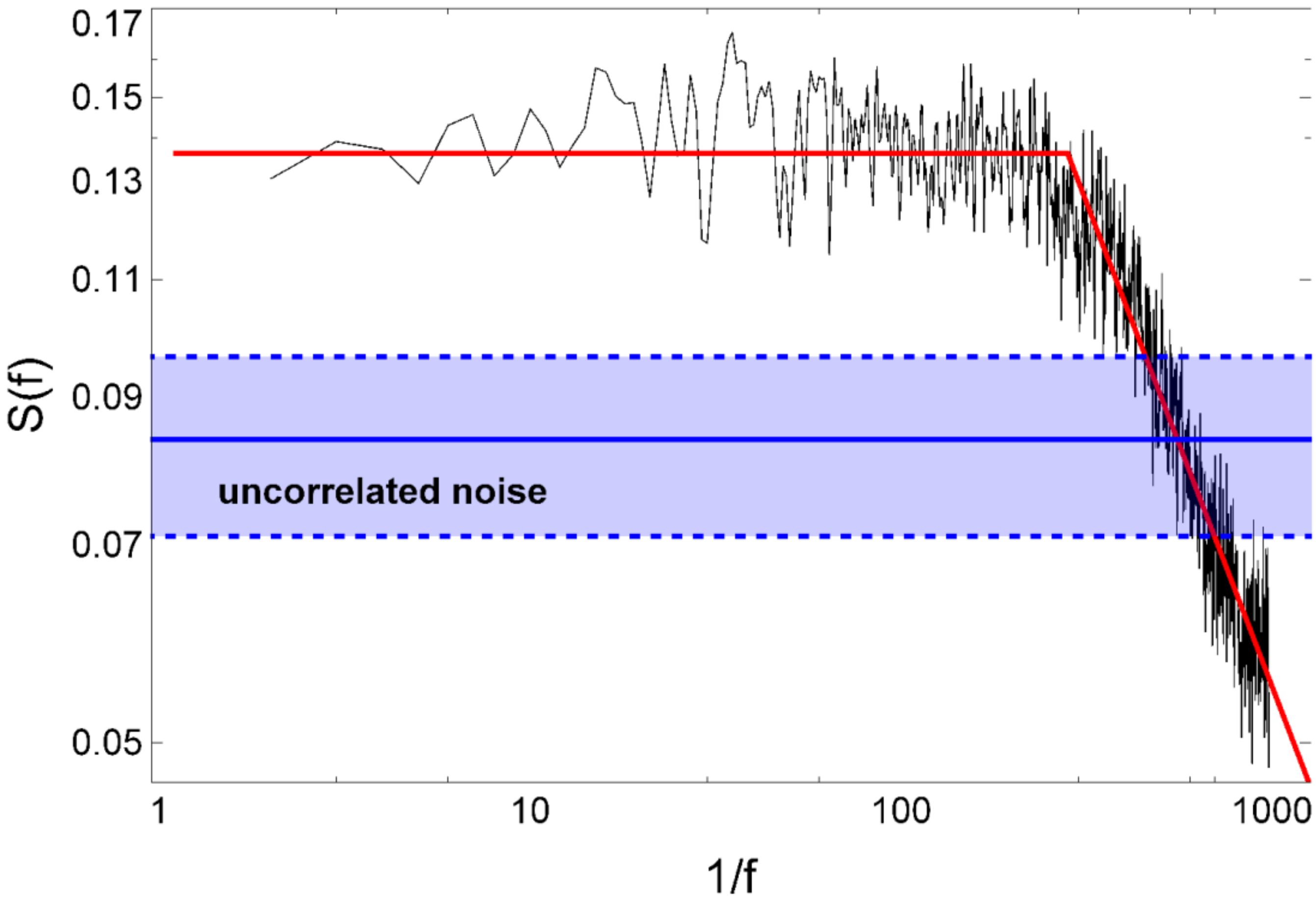




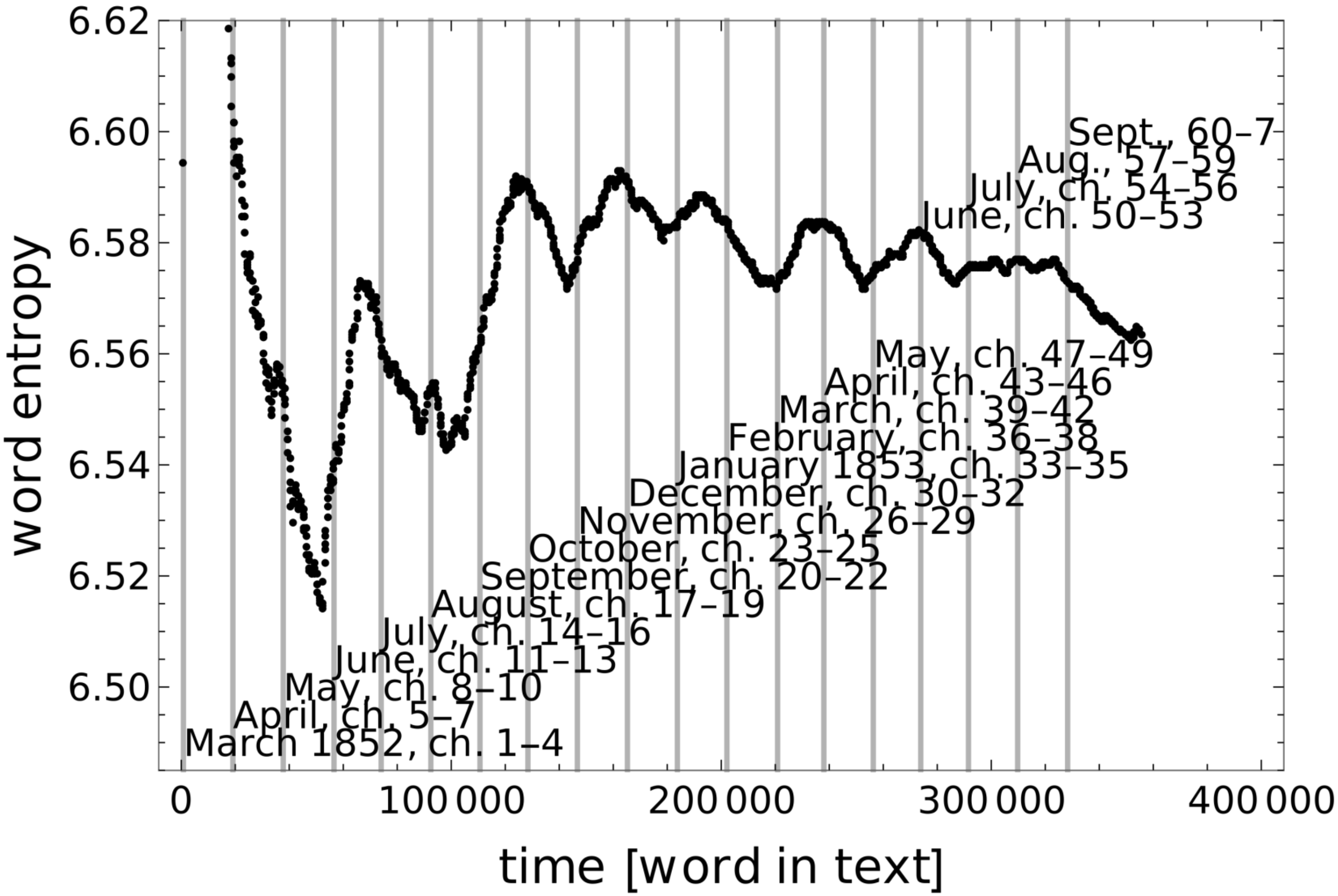
Źródło: https://en.wikipedia.org/wiki/Spectral_density

Power spectrum of time series

$$S(f) = |\sum_t x_t \exp(-2\pi i f t)|^2 \qquad (1)$$



Development of vocabulary



Ochab JK (2016). Time Series Analysis Enhances Authorship Attribution. *Digital Humanities* conference abstracts, July 11-16, 2016, Kraków.

Take-home message:

- „Temporal" features can be used to characterise and classify texts, too.

A. Pawłowski, in Travaux de linguistique quantitative, Vol. 62 (Honoré Champion, Paris, Geneve: Champion-Slatkine, 1998).
A. Pawłowski, Journal of Quantitative Linguistics 6, 70 (2011).

Conclusions

- Texts can be quantified in a number of ways
- Technological breakthrough not only for tech companies but for research in humanities

Based on:

Ochab JK, Byszuk J, Pielström S, Eder M (2019)
Identifying Similarities in Text Analysis: Hierarchical Clustering (Linkage) versus Network Clustering (Community Detection).
Digital Humanities conference abstracts, July 9-12, 2019, Utrecht.

Škvrňák J, Škvrňák M, Ochab JK (2019)
How To Detect Coup d'État 800 Years Later.
Digital Humanities conference abstracts, July 9-12, 2019, Utrecht.

Ochab JK, Essler H (2019)
Stylometry of literary papyri.
3rd International Conference on Digital Access to Textual Cultural Heritage (DATeCH2019), May 8-10, 2019, Brussels, Belgium.

Franzini G, et al. (2018)
Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm.
Front. Digit. Humanit. 5:4

Ochab JK (2017)
Stylometric networks and fake authorships.
Leonardo 50

Ochab JK (2017)
Randall Munroe's Thing Explainer: The Tasks in Translation of a Book Which Explains the World With Images.
Przekładaniec 34-35

Ochab JK (2016)
Time Series Analysis Enhances Authorship Attribution.
Digital Humanities conference abstracts, July 11-16, 2016, Kraków.

Universiteit Antwerpen

M Kestemont

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

DECODED

M Büchler
G Franzini
G Rotari
M Jander
E Franzini

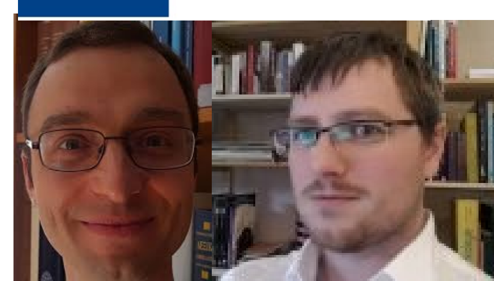
PAN IJP Institute of Polish Language Polish Academy of Sciences

Institute of English Studies
Jagiellonian University

Julius-Maximilians-UNIVERSITÄT WÜRZBURG

M Eder
J Byszuk

J Rybicki



H Essler
S Pielström

10 Computational 01
01 Stylistics 0101000
11 Group 011010110

computationalstylistics.github.io

github.com/computationalstylistics/

dh
K

facebook.com/dhkrakow/

Digital Humanities Lab

Flagship Project at Jagiellonian University

https://dhlab.id.uj.edu.pl/