

Deep learning in the humanities



Methodological Department
at the Institute of Polish Language PAN
Maciej Eder, Michał Woźniak, Joanna Byszuk

10 Computational 01
01 Stylistics 0101000
11 Group 011010110

PAN JP

A few applications



- OCR/HTR – optical character recognition / handwritten text recognition
- Modeling language and its various aspects,
 - Tracing language change over time
- Classification – stylistic analysis, authorship attribution
- Interdisciplinary research:
 - The language of chemistry

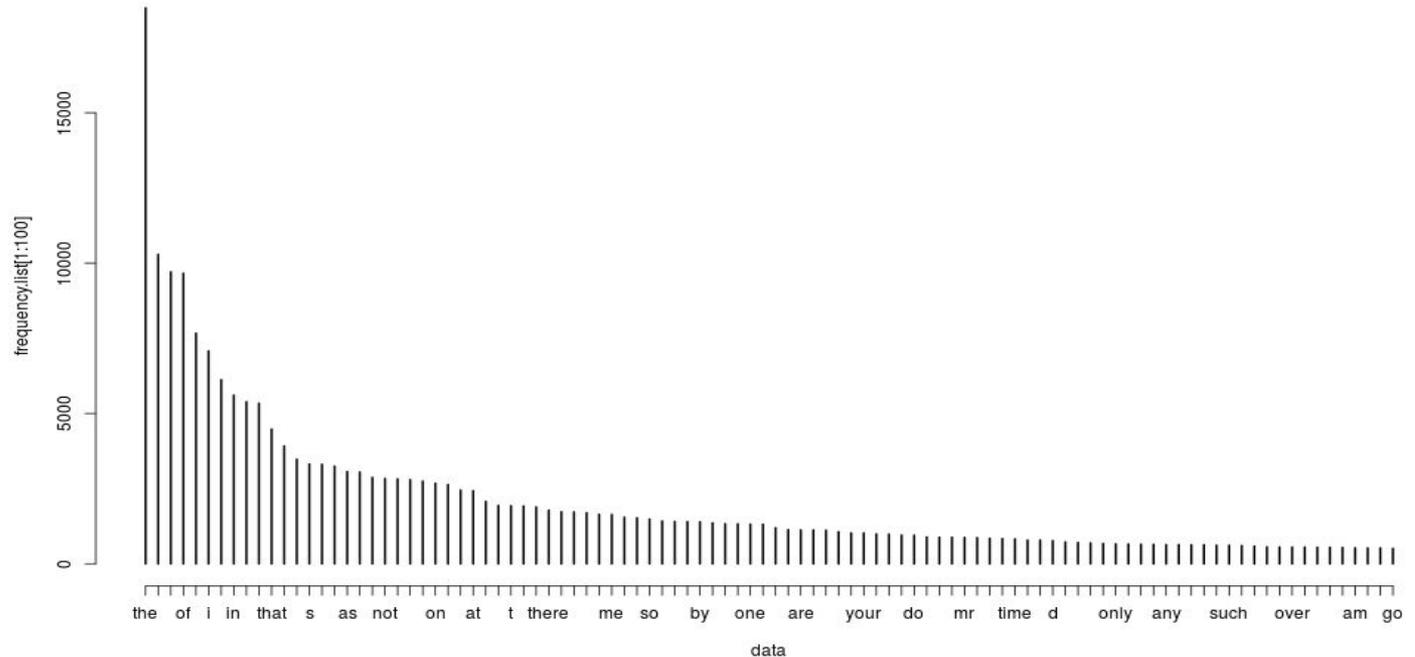
Stylometry and authorship attribution

Stylometry =

**use of quantitative methods
to examine similarities and
differences within a group of
texts**



Stylometry and authorship attribution



The possibility of using **frequency patterns of very common words** rests upon the fact that words do not function as discrete entities. Since they gain their full meaning through the different sorts of **relationship they form with each other**, they can be seen as **markers of those relationships** and, accordingly, of everything that those relationships entail.

Stylometry and authorship attribution – applications



- authorship attribution,
- tracing chronology,
- analysis of cross and inter genre relationships,
- big data analysis,
- style transfer and anonymization,
- and many others.

Stylometry and authorship attribution – how to



corpus of texts

+

distance measure

+

classification algorithm

+

(visualisation)

Stylometry and authorship attribution – how to



	Agnes	Tenant	Emma	Pride	Sense	Jane
the	2511	5929	5204	4330	4105	7835
and	2733	6705	4878	3577	3489	6618
to	2366	5594	5186	4136	4103	5152
of	1602	3734	4292	3609	3571	4359
i	2204	6075	3191	2064	1998	7165
a	1296	2792	3126	1948	2067	4467
in	911	2021	2174	1866	1948	2762
that	776	1909	1800	1577	1383	1655
he	659	2259	1811	1338	1112	1902
was	1000	1835	2400	1847	1861	2525
it	795	2280	2529	1532	1755	2403
you	760	2844	1999	1356	1191	2971
her	750	1760	2483	2224	2543	1714

Stylometry and authorship attribution – how to



	Agnes	Tenant	Emma	Pride	Sense	Jane
the	3.67471	3.54285	3.24344	3.55705	3.43227	4.18704
and	3.99959	4.00655	3.04026	2.93847	2.91722	3.53667
to	3.46251	3.34267	3.23222	3.39768	3.43060	2.75324
of	2.34444	2.23124	2.67503	2.96476	2.98579	2.32946
i	3.22543	3.63009	1.98882	1.69556	1.67057	3.82899
a	1.89662	1.66835	1.94831	1.60026	1.72826	2.38717
in	1.33320	1.20764	1.35496	1.53290	1.62876	1.47602
that	1.13563	1.14072	1.12187	1.29549	1.15635	0.88444
he	0.96441	1.34986	1.12872	1.09915	0.92977	1.01643
was	1.46344	1.09650	1.49582	1.51729	1.55602	1.34937
it	1.16344	1.36241	1.57622	1.25852	1.46739	1.28417
you	1.11222	1.69942	1.24589	1.11394	0.99582	1.58771
her	1.09758	1.05168	1.54755	1.82699	2.12625	0.91597

Stylometry and authorship attribution – how to

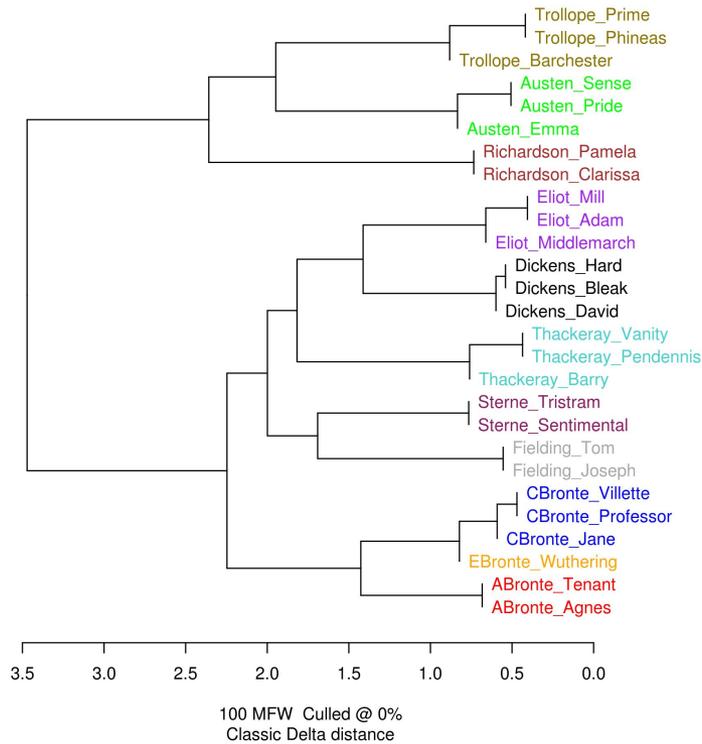


	Agnes	Pride	Jane	David	Mill	Tom	Clarissa
Tenant	0.81	1.07	0.88	0.92	0.98	1.16	1.1
Emma	1.12	0.78	1.28	1.15	1.2	1.25	1.24
Sense	1.14	0.69	1.24	1.16	1.25	1.13	1.21
Professor	1.06	1.21	0.69	0.94	1	1.27	1.3
Villette	1.07	1.26	0.65	0.91	0.96	1.28	1.3
Bleak	1.09	1.18	0.92	0.55	0.87	1.21	1.17
Hard	1.16	1.25	0.96	0.65	0.91	1.26	1.25
Wuthering	1.06	1.31	0.81	0.94	1.01	1.32	1.27
Adam	1.13	1.37	0.95	0.9	0.66	1.42	1.32
Middlemarch	1.01	1.1	0.99	0.87	0.65	1.17	1.12
Joseph	1.2	1.19	1.24	1.18	1.29	0.64	1.11
Pamela	1.15	1.24	1.27	1.19	1.26	1.11	0.67
Sentimental	1.38	1.53	1.23	1.22	1.29	1.42	1.38

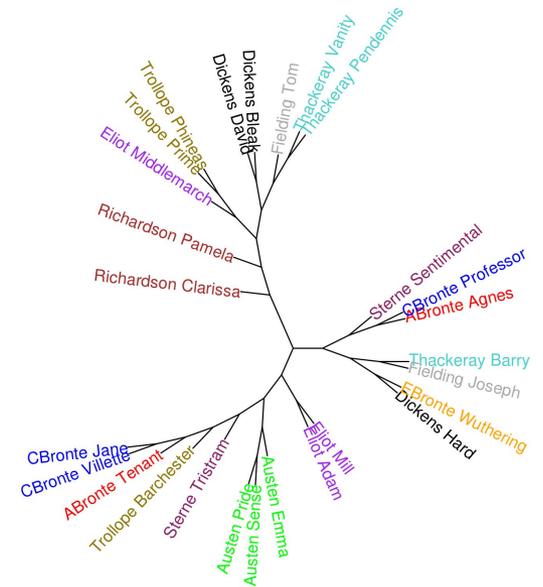
Stylometry and authorship attribution – how to



**ASmallCollection
Cluster Analysis**



**ASmallCollection
Bootstrap Consensus Tree**



Optical Character Recognition & Handwritten Text Recognition

OCR / HTR

- Turning large text collections into digital corpora
- Traditional OCR/HTR approaches - **7% to 85% accuracy**

Z a c h ę c o n y dobrem przyięciem G ram m aty k i llo ssy y sk iey , k tó
rey drugie w y danie w ro k u 181 i nastąpiło; dla przysługi W spótzioinkóm ,
spieszę się z ofiarow aniem Polskiej, w tym zamiarze ,
ażeby ona do popraw y błędnego dziś
i (śmiało rzec m ożna) nadto przesadzon ego, a ze zwyczajem praw dziw ie
Polskim niezgodnego pisania i m ów ienia
sp o so b u , w czym kolw iek dopomódz
m ogła.

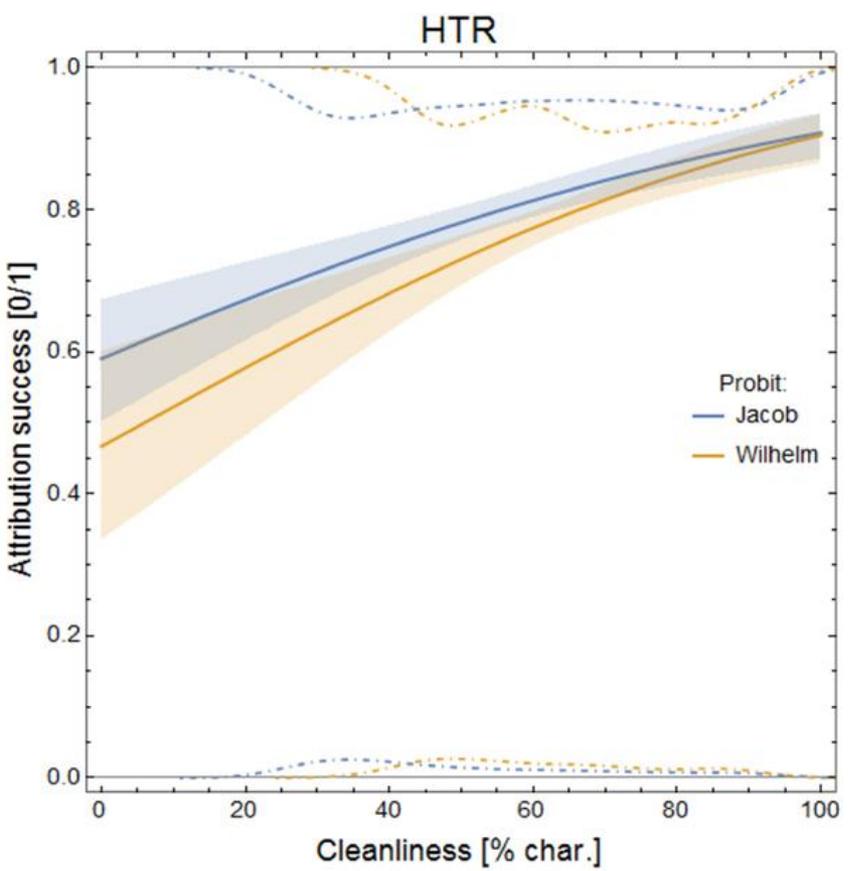
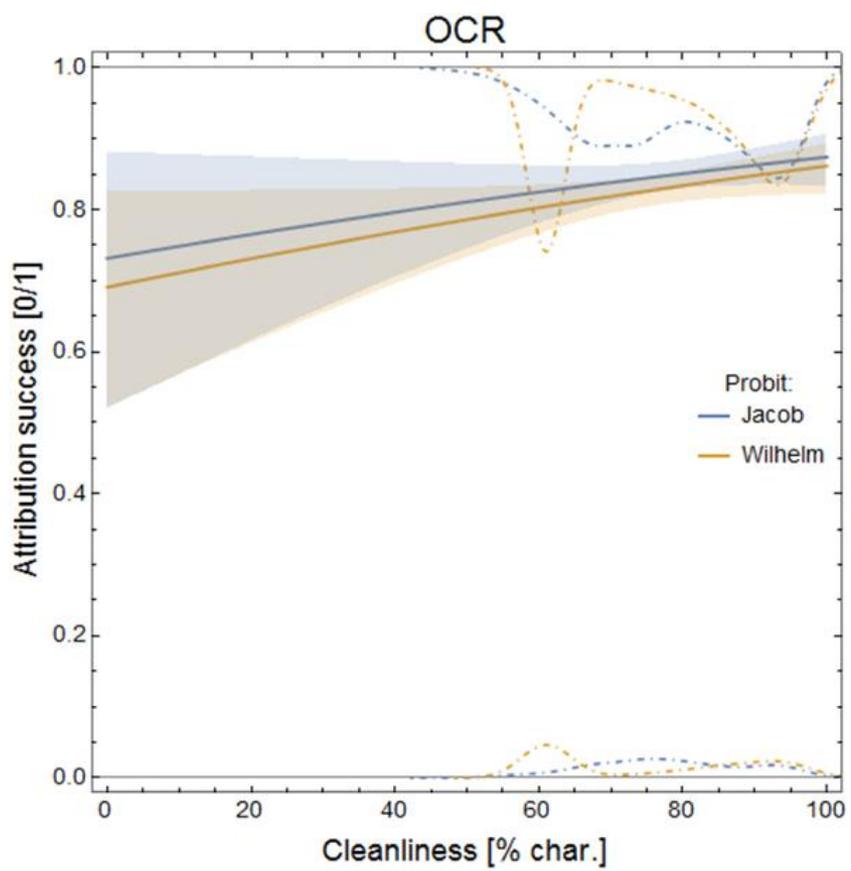
Nie miesce tu w y ty k a ć , iakie się
W pisaniu i ustney m owie zagęściły błęd y : przez iednych um yślnie (że tak
pow iem) śpiknionych na zniszczenie piękności ięzy k a, i dla zaprowadzenia,
podpozorem now om odnych a dziw acznych
w y razów , zawilóści rzeczy: przez d ru gich, z niewiadomości p raw ideł
popelniane: ale ktoby życzył u niknąć zdrożności; w tern sozupłem dziełku
znaydzie niektóre podane sobie rady.

Może mi kto mieć za złe i zarzucić, zem te praw id ła s cudzey w y p isał p
race: nie przeczę ia te m u : bom się
sczerze chciał w e wszystkim stosować
ta k do woli niegdyś za królestw a Poln

. ■ ■ .

OCR / HTR

- Works for rough classification
- Not enough for studies on variation



OCR / HTR

- Neural-based approaches – with enough training data much greater accuracy

The screenshot displays a handwriting recognition interface. On the left, there is a legend for various text elements (add, blackening, date, div, aao, organization, person, place, sic, signature, speech, suoolled, textStvie, unclear, work) with corresponding color swatches. Below the legend, there are fields for 'Tags under cursor' and 'Properties of 'person' tag'. The main area shows a handwritten letter with numbered annotations (14-21) and a corresponding transcription table at the bottom. The signature 'Jac. Grimm' is also visible.

Property	Value
offset	0
lenath	10
continued	false
notice	
occupation	
firstname	Lamprecht
dateOfDeath	nach 1250
dateOfBirth	1215
lastname	von Regensburg

14	handlungen scheint es gar nicht.
15	Lamprechts tochter Syon verdient sicherlich eine ausgabe, und
16	in ermangelung bequemerer verleger steht die Quedlinburger
17	nationalbibliothek dafür offen. Basse gewährt auch andstän-
18	dige honorare.
19	Mich hochachtungsvoll empfehend
20	Jac. Grimm
21	Cassel 29 mai 1840

Modeling linguistic change

Modeling and Visualizing the Dynamics of Linguistic Changes

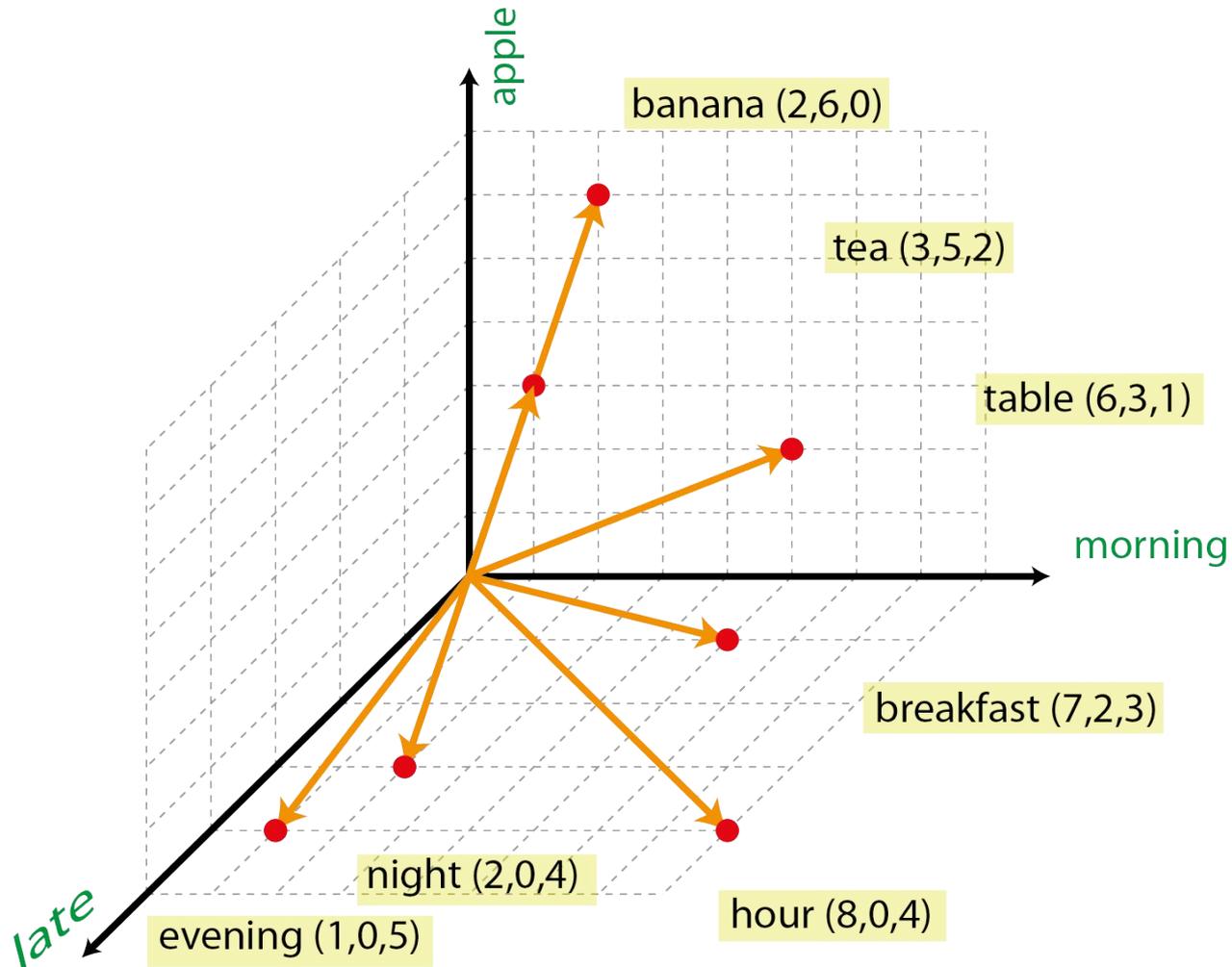
Problem:

Vast collections of data – can we trace evolution of the language?

Solution:

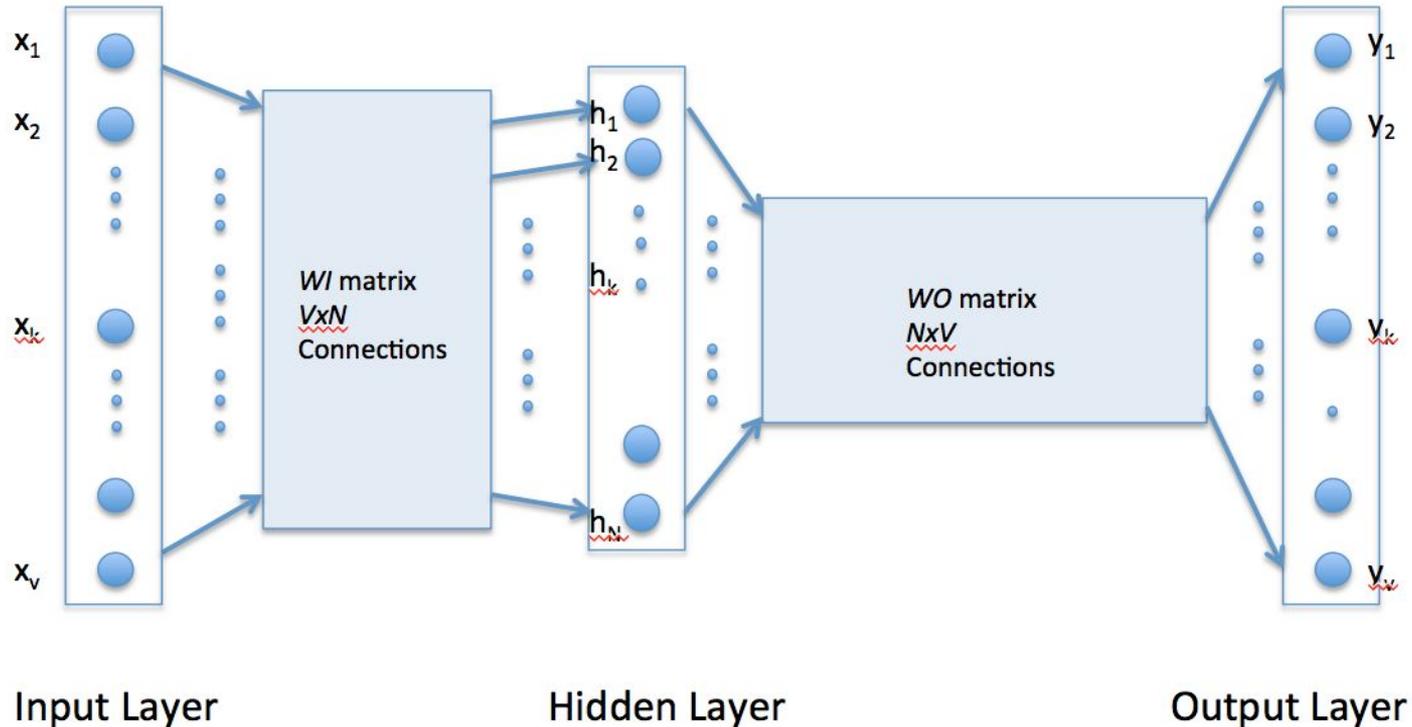
- tracking change with logistic regression
- comparing word vectors over the years

Modeling and Visualizing the Dynamics of Linguistic Changes

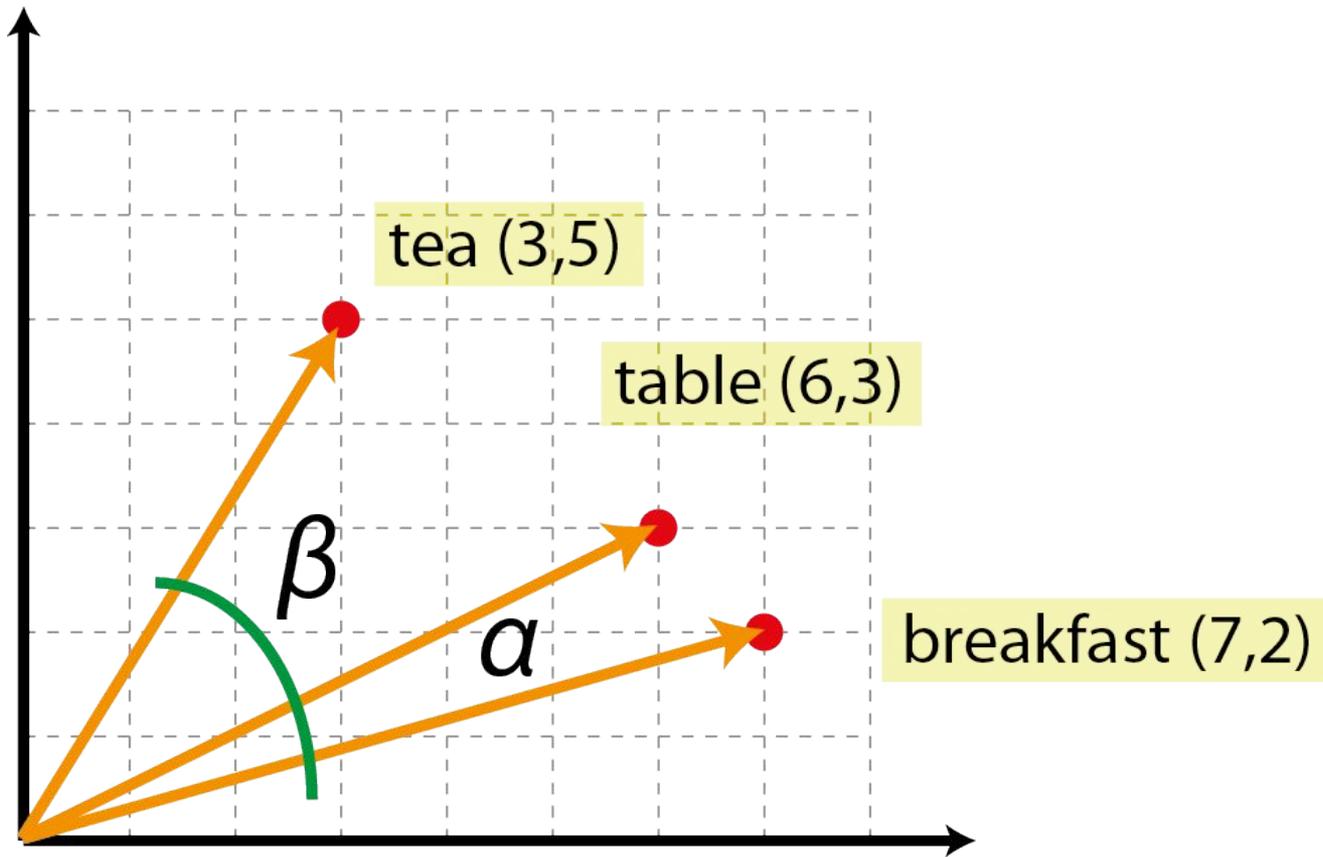


Modeling and Visualizing the Dynamics of Linguistic Changes

word2vec



Modeling and Visualizing the Dynamics of Linguistic Changes



Modeling and Visualizing the Dynamics of Linguistic Changes

'man' – 'woman' + 'trousers' = 'shirt'

```
word_embedding_models — R — 80x24
>
>
>
>
> my_vector = word_vectors["man", , drop = FALSE] - word_vectors["woman", , drop
= FALSE] + word_vectors["trousers", , drop = FALSE]
> cos_sim = sim2(x = word_vectors, y = my_vector, method = "cosine", norm = "l2")
> head(sort(cos_sim[,1], decreasing = TRUE), 25)
trousers      shirt  flannel      coat      belt  throats  drained  pockets
0.8485228 0.7142671 0.6774317 0.6569998 0.6271524 0.6206532 0.6107239 0.6102036
sleeves      tail waistcoat breeches  tails  boots    bone    collar
0.5964580 0.5889789 0.5862612 0.5842724 0.5652466 0.5627244 0.5539768 0.5519815
sleeve      horse mustache  rubbed    wig     vote    cravat   stuck
0.5514129 0.5419616 0.5389607 0.5386861 0.5279935 0.5270436 0.5238806 0.5232810
steel
0.5222099
>
>
>
>
>
>
>
```

Chemical linguistics

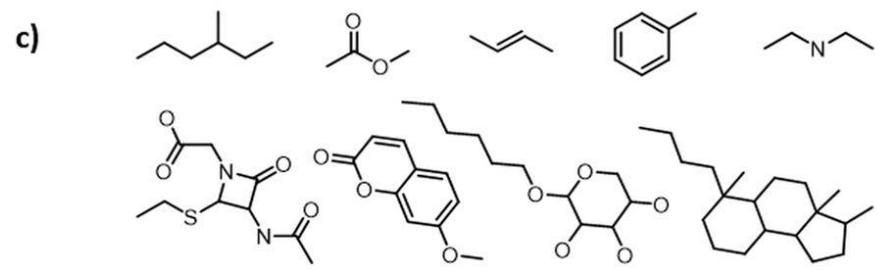
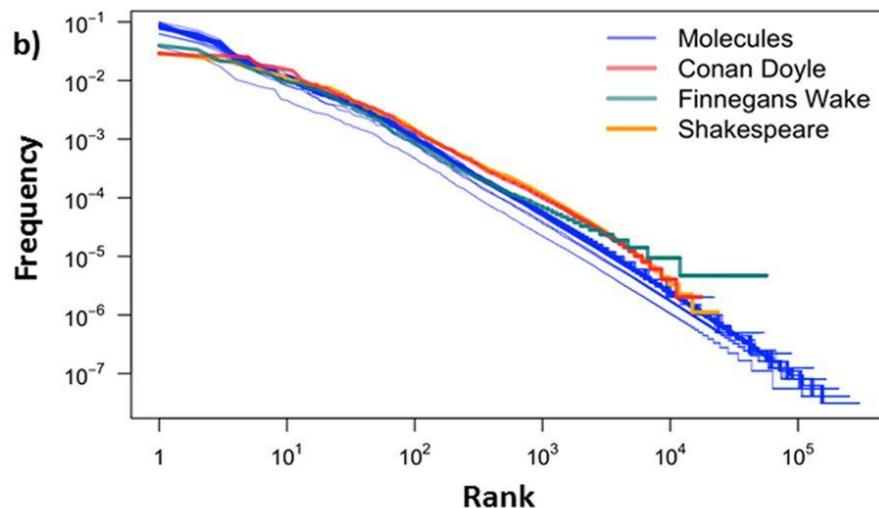
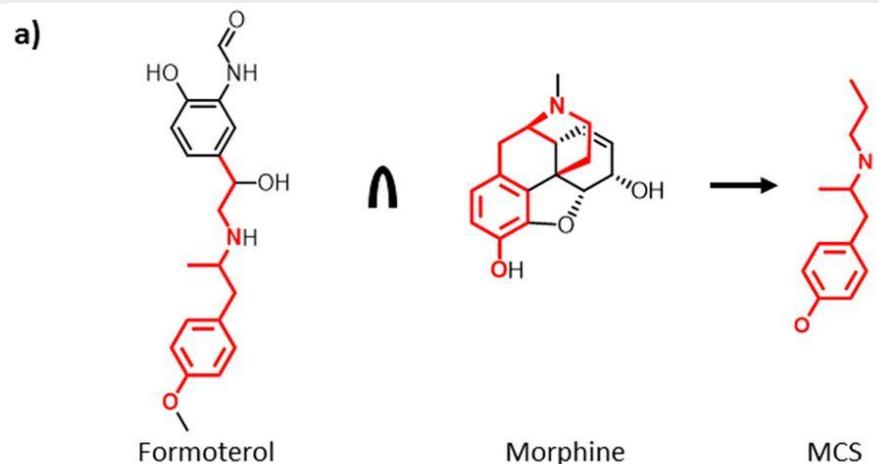
Chemical linguistics

Chemical words and vocabularies.

(a) A common maximal substructure, MCS (colored red), between two molecules.

(b) Language laws in MSC “words” for the entire 1.75-million-rich chemical vocabulary, compared with the works by Conan Doyle, Shakespeare, and Joyce’s Finnegans Wake.

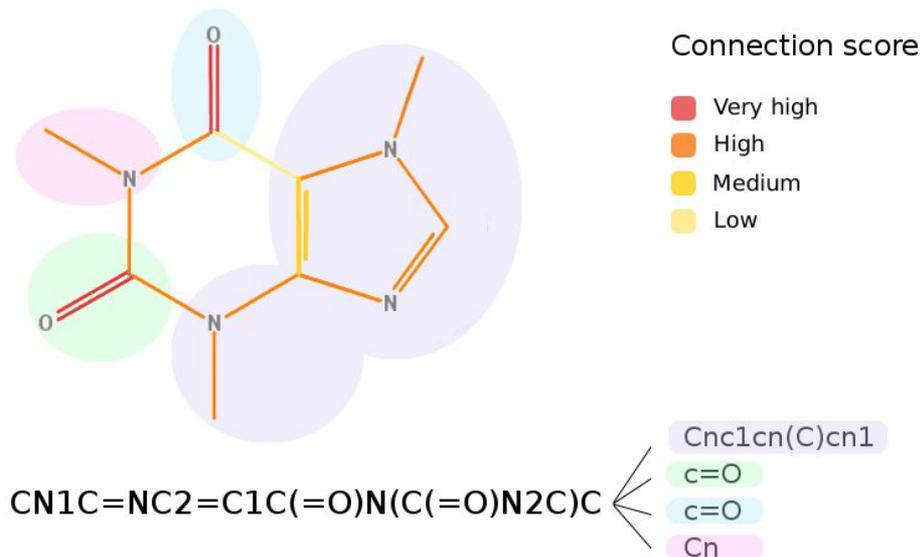
(c) Examples of chemical words – frequent “function” versus infrequent “content” words (from left to right: penicillins, coumarins, carbohydrates, steroids).



Chemical linguistics

Chemical words – second approach.

- Molecule as a graph
- *Mutual Information* indicates strength of bonds between atoms
- Groups (communities) detected by *Louvain algorithm*





Chemical linguistics

Basic deep learning classification

Three chemical databases:

- General (NOC)
- Drugs
- Illegal substances

descriptor	drugs vs NOC	illegal vs NOC
functional groups	73.46%	71.77%
fingerprints	77.46%	79.62%
mcs	75.78%	81.6%
communities	78.89%	82.59%



More about our projects

<https://computationalstylistics.github.io/projects/>

Maciej Eder

Michał Woźniak

@ijp.pan.pl

Joanna Byszuk