



Quantitative Textanalyse 1: Stilometrie

Vorlesung *Einführung in die Digital Humanities*
MSc Digital Humanities | Wintersemester 2020/21

Prof. Dr. Christof Schöch



Sitzungsüberblick

1. Quantitative Textanalyse: Überblick
2. Software im Überblick
3. Stilometrie (Textähnlichkeit)

(1) Quantitative Textanalyse: Überblick

Anwendungsbereiche

3.1

- Autorschaftsattributions
- Gattungsstilistik
- Netzwerkanalyse
- Inhaltsanalyse (Begriffe, Topics)
- Automatische Kartierung
- Extraktion von Zeitstrukturen
- Erkennung erzähltheoretischer Kategorien
- uvm.

3.2

Grundlegende Verfahren

- Suche nach Mustern
- Kontrastive Analyse
- Gruppen ähnlicher Texte entdecken
- Verteilungen und Entwicklungen finden
- Informationen explizit machen
- Dimensionsreduktion
- Texte klassifizieren

Perspektiven der digitalen Textanalyse

- Quantitative vs. qualitative Verfahren
- Analyseverfahren vs. Datenvisualisierung
- Grafische Benutzungsoberfläche (GUI) vs. Command Line Interface (CLI)
- Klassifikation vs. Clustering

(2) Software im Überblick

Natural Language Processing: NLTK

4.1

1 Using a Tagger

A part-of-speech tagger, or **POS-tagger**, processes a sequence of words, and attaches a part of speech tag to each word (don't forget to `import nltk`):

```
>>> text = word_tokenize("And now for something completely different")
>>> nltk.pos_tag(text)
[('And', 'CC'), ('now', 'RB'), ('for', 'IN'), ('something', 'NN'),
 ('completely', 'RB'), ('different', 'JJ')]
```

Here we see that *and* is `cc`, a coordinating conjunction; *now* and *completely* are `RB`, or adverbs; *for* is `IN`, a preposition; *something* is `NN`, a noun; and *different* is `JJ`, an adjective.

Note

NLTK provides documentation for each tag, which can be queried using the tag, e.g. `nltk.help.upenn_tagset('RB')`, or a regular expression, e.g. `nltk.help.upenn_tagset('NN.*')`. Some corpora have README files with tagset documentation, see `nltk.corpus.??? readme()`, substituting in the name of the corpus.

Let's look at another example, this time including some homonyms:

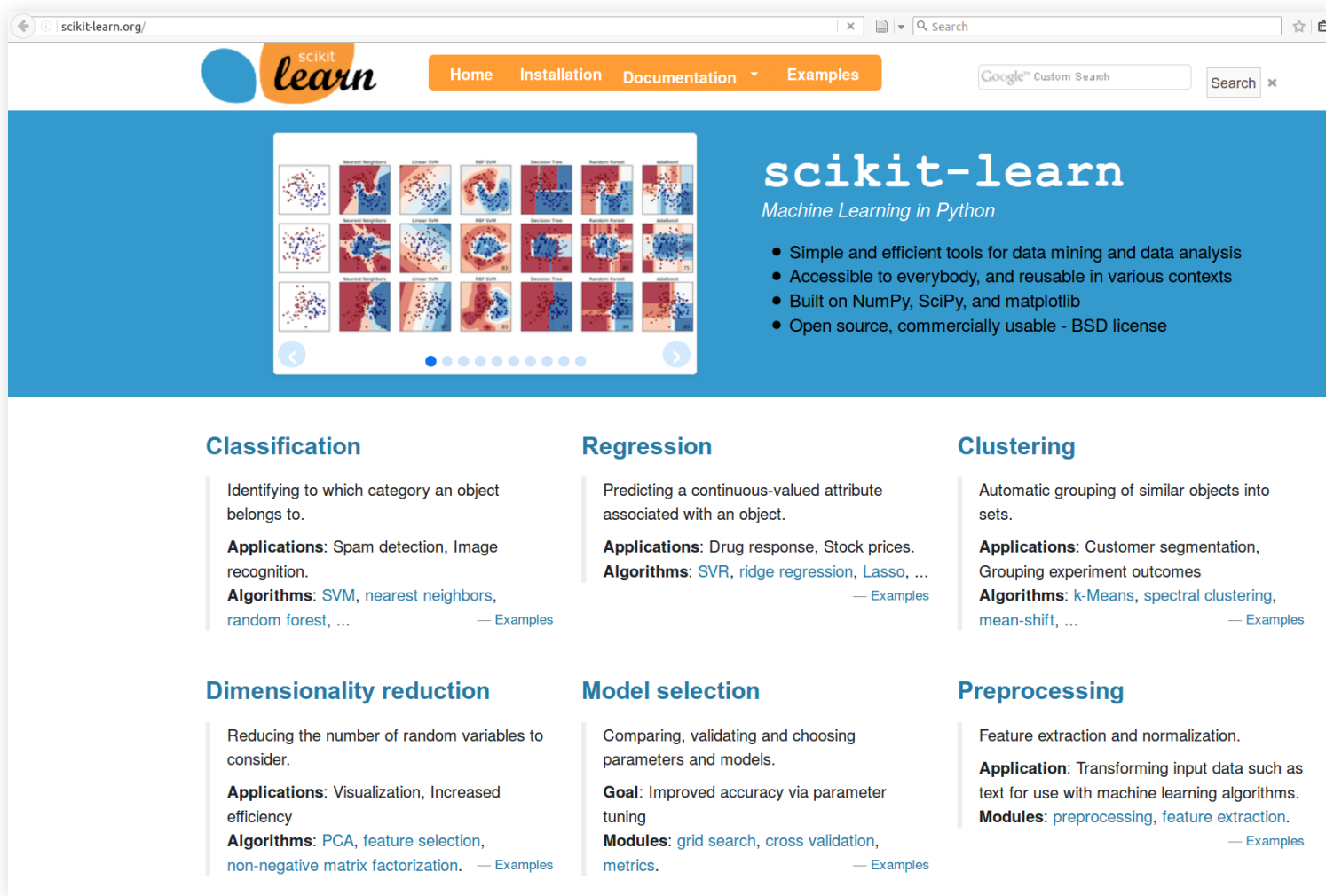
```
>>> text = word_tokenize("They refuse to permit us to obtain the refuse permit")
>>> nltk.pos_tag(text)
[('They', 'PRP'), ('refuse', 'VBP'), ('to', 'TO'), ('permit', 'VB'), ('us', 'PRP'),
 ('to', 'TO'), ('obtain', 'VB'), ('the', 'DT'), ('refuse', 'NN'), ('permit', 'NN')]
```

Notice that *refuse* and *permit* both appear as a present tense verb (`VBP`) and a noun (`NN`). E.g. *refUSE* is a verb meaning "deny," while *REFuse* is a noun meaning "trash" (i.e. they are not homophones). Thus, we need to know which word is being used in order to pronounce the text correctly. (For this reason, text-to-speech systems usually perform POS-tagging.)

<http://www.nltk.org>; Alternativen: TreeTagger, spaCy

4.2

Maschinelles Lernen: scikit-learn (Python)



The screenshot shows the scikit-learn website in a web browser. The browser's address bar displays 'scikit-learn.org/'. The website's header includes the 'scikit-learn' logo, navigation links for 'Home', 'Installation', 'Documentation', and 'Examples', and a Google Custom Search bar. The main banner features a 3x3 grid of 12 small plots illustrating various machine learning models, with the text 'scikit-learn Machine Learning in Python' and a list of key features: 'Simple and efficient tools for data mining and data analysis', 'Accessible to everybody, and reusable in various contexts', 'Built on NumPy, SciPy, and matplotlib', and 'Open source, commercially usable - BSD license'. Below the banner, six categories are listed: Classification, Regression, Clustering, Dimensionality reduction, Model selection, and Preprocessing. Each category includes a brief description, applications, algorithms, and a link to examples.

Classification
Identifying to which category an object belongs to.
Applications: Spam detection, Image recognition.
Algorithms: SVM, nearest neighbors, random forest, ... — Examples

Regression
Predicting a continuous-valued attribute associated with an object.
Applications: Drug response, Stock prices.
Algorithms: SVR, ridge regression, Lasso, ... — Examples

Clustering
Automatic grouping of similar objects into sets.
Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, ... — Examples

Dimensionality reduction
Reducing the number of random variables to consider.
Applications: Visualization, Increased efficiency
Algorithms: PCA, feature selection, non-negative matrix factorization. — Examples

Model selection
Comparing, validating and choosing parameters and models.
Goal: Improved accuracy via parameter tuning
Modules: grid search, cross validation, metrics. — Examples

Preprocessing
Feature extraction and normalization.
Application: Transforming input data such as text for use with machine learning algorithms.
Modules: preprocessing, feature extraction. — Examples

<http://scikit-learn.org>, Alternative: R

Korpusanalyse: TXM

The screenshot displays the TXM software interface with the following components:

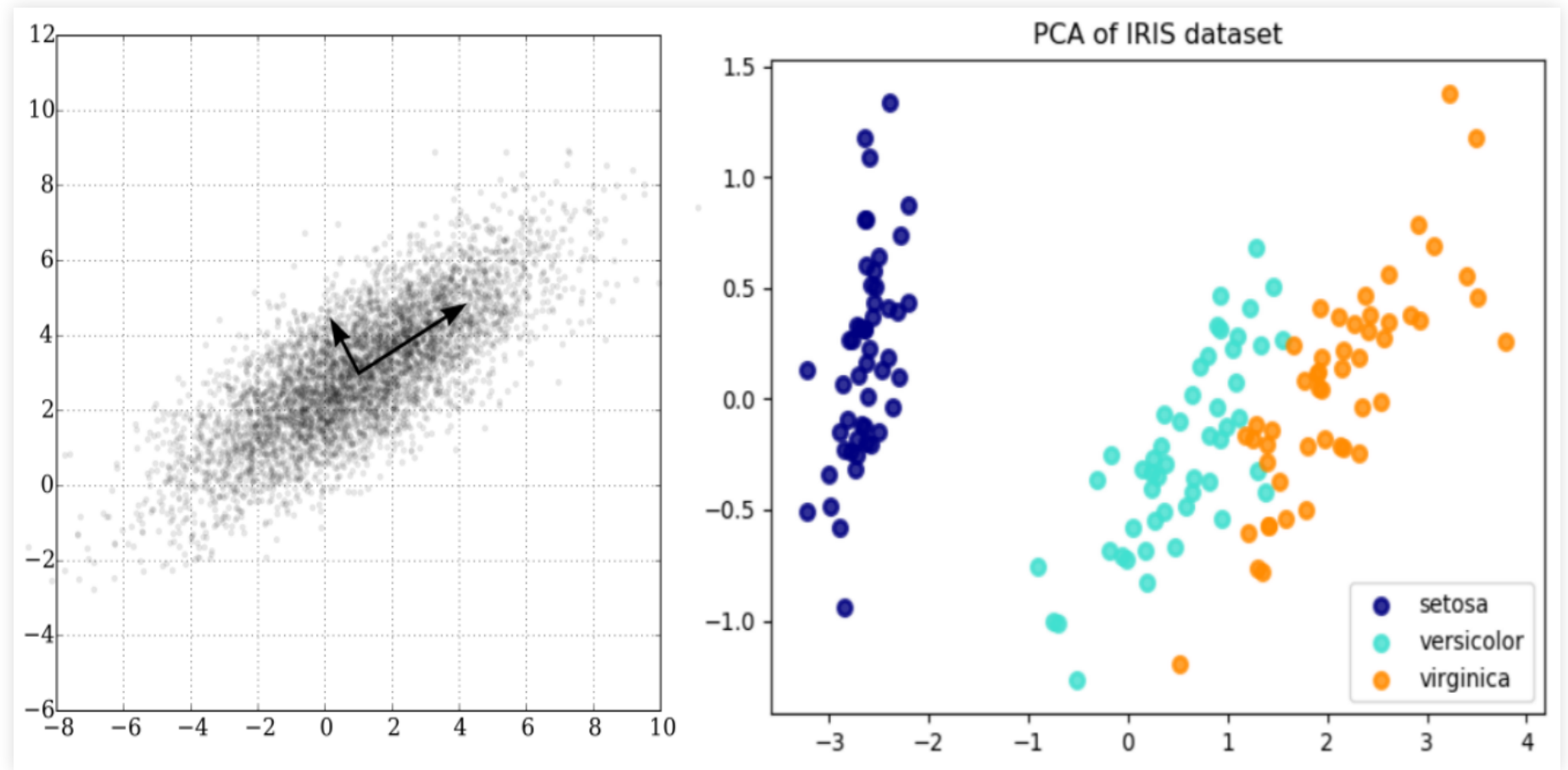
- Corpus List (Left):** Shows a tree view of corpora including GRAAL, MANCHETTE, TC377, DOYLE, TXMFINAL, and ROMAN19. The selected corpus is ROMAN19, with a subquery for `[frlemma="musique"] (5, 5)`.
- Query Panel (Top Left):**
 - Query: `[frlemma="musique"]`
 - Cooccurrences properties: word (selected), structure (disabled)
 - Thresholds: Fmin ≥ 2, Cmin ≥ 2, Score ≥ 4
 - Context: word (selected), structure (disabled)
 - Options: ☒ use left Window, ☒ use right Window, ☐ include the keyword structure
 - Range: from -5 to 0 and from 0 to 5
- Results Table (Bottom Left):**

Cooccurrent	Frequency	Cofrequency	Score	Mean distance
la	99125	161	34	.7
de	172450	162	11	1.4
instruments	130	5	7	1.6
poésie	88	4	6	3.0
instrumentale	2	2	6	2.0
ophicléides	3	2	6	3.0
Théorie	3	2	6	2.0
militaire	138	4	5	.0
entendre	1136	7	5	3.1
copier	6	2	5	1.5
sons	58	3	5	2.3
musique	221	4	5	2.5
régiments	11	2	4	1.0
relier	12	2	4	1.0
pour	24246	30	4	2.2
maître	2145	8	4	1.6
des	32495	35	4	2.6
Nicétas	164	3	4	4.0
jouer	445	4	4	2.5
- Query Panel (Top Right):**
 - Query: `[word="instruments|poésie"] []* [frlemma="musique"] within 6`
 - Keyword: word (selected)
 - Sort keys: #1 None, #2 None, #3 None, #4 None, Sort
 - Navigation: < 1 -9/9 >
 - Hide settings button
- Context View (Bottom Right):**

Left context	Keyword	Right context
principe vital Est une hypothèse gratuite ! La	poésie et la musique	parurent admirables à L
marchand de vin, par un facteur d'	instruments de musique	et par un libraire qui ver
et des pincettes, des manchons et des	instruments de musique	, des dentelles, des mari
a des festivals pour tout : pour la	musique, pour la poésie	, pour le vin blanc, pour l
le mur, étaient posés ou accrochés des	instruments de musique	, des violons, des cornet
des paysans couvrait entièrement parfois la chanson des	instruments ; et la frêle musique	déchirée par les voix dé
et, d'ailleurs, assez frottée de	musique, de peinture et de poésie	, depuis son enfance, po
sait toutes choses mieux que moi : la	poésie, la musique	, les langues, les science
, vos livres, vos chasses, vos	instruments de musique	vous attendent, les être
- Console (Bottom):**
 - System Output
 - Found 9 occurrences

<http://textometrie.ens-lyon.fr>; Alternative: Antconc

Principal Components Analysis



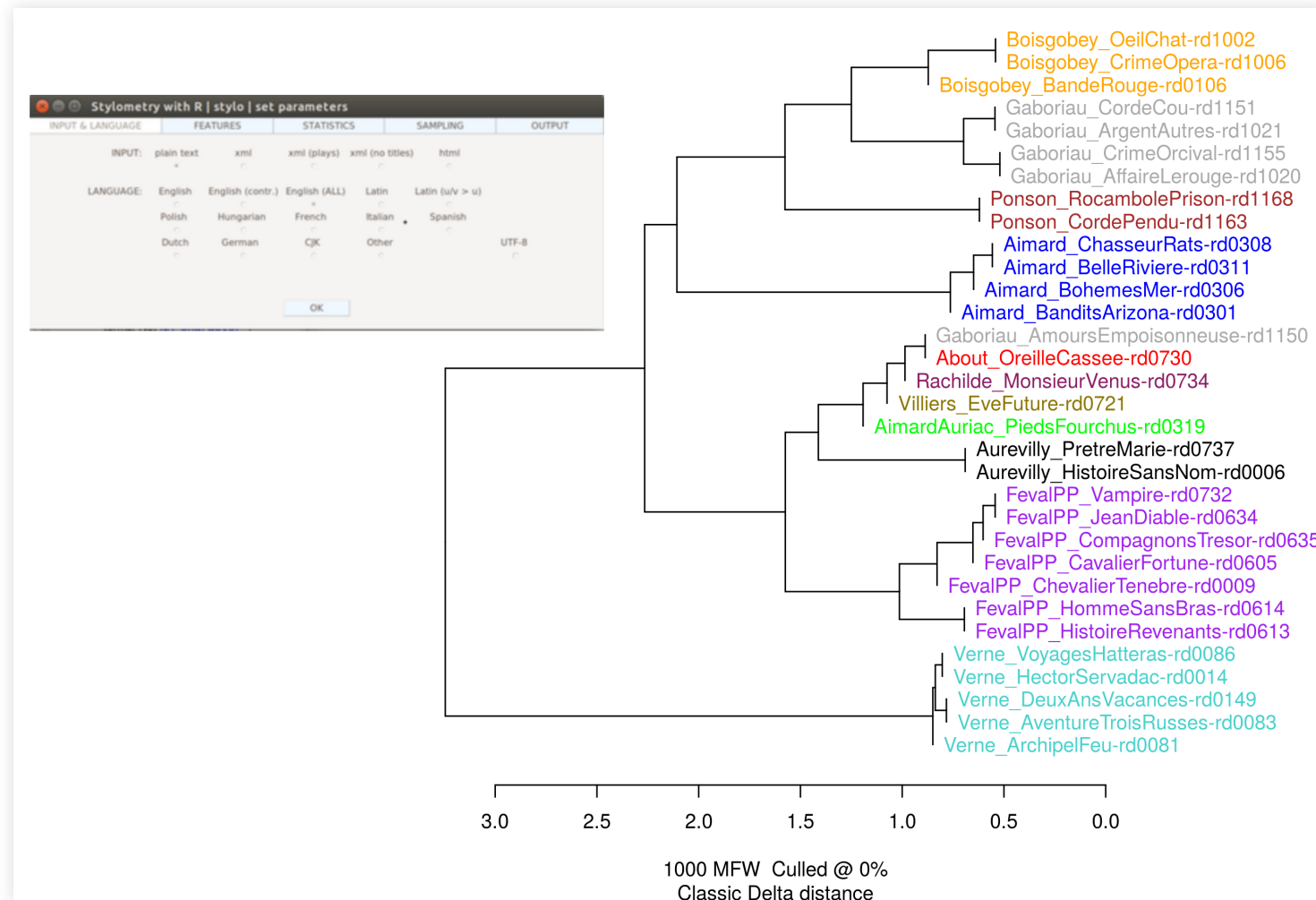
Diverse Tools: stylo, Python, etc.

Topic Modeling: MALLET

```
christof@DELL: ~/Programs/mallet
christof@DELL:~$ cd Programs/mallet
christof@DELL:~/Programs/mallet$ bin/mallet import-dir --input /home/christof/Repos/clgs/polar/txt10d --output polar.mallet --keep-sequence --token-regex '\p{L}[\p{L}\p{P}]*\p{L}' --remove-stopwords TRUE --stoplist-file stoplists/fr3.txt
Labels =
    /home/christof/Repos/clgs/polar/txt10d
christof@DELL:~/Programs/mallet$ bin/mallet train-topics --input polar.mallet -
-num-topics 30 --optimize-interval 200 --num-iterations 4000 --num-top-words 30
--word-topic-counts-file results/polar10d_words-by-topics.txt --output-state top
ic-state.gz --output-topic-keys results/polar-10d_topics-with-words.txt --output
-doc-topics results/polar10d_topics-in-texts.txt --doc-topics-max 30
```

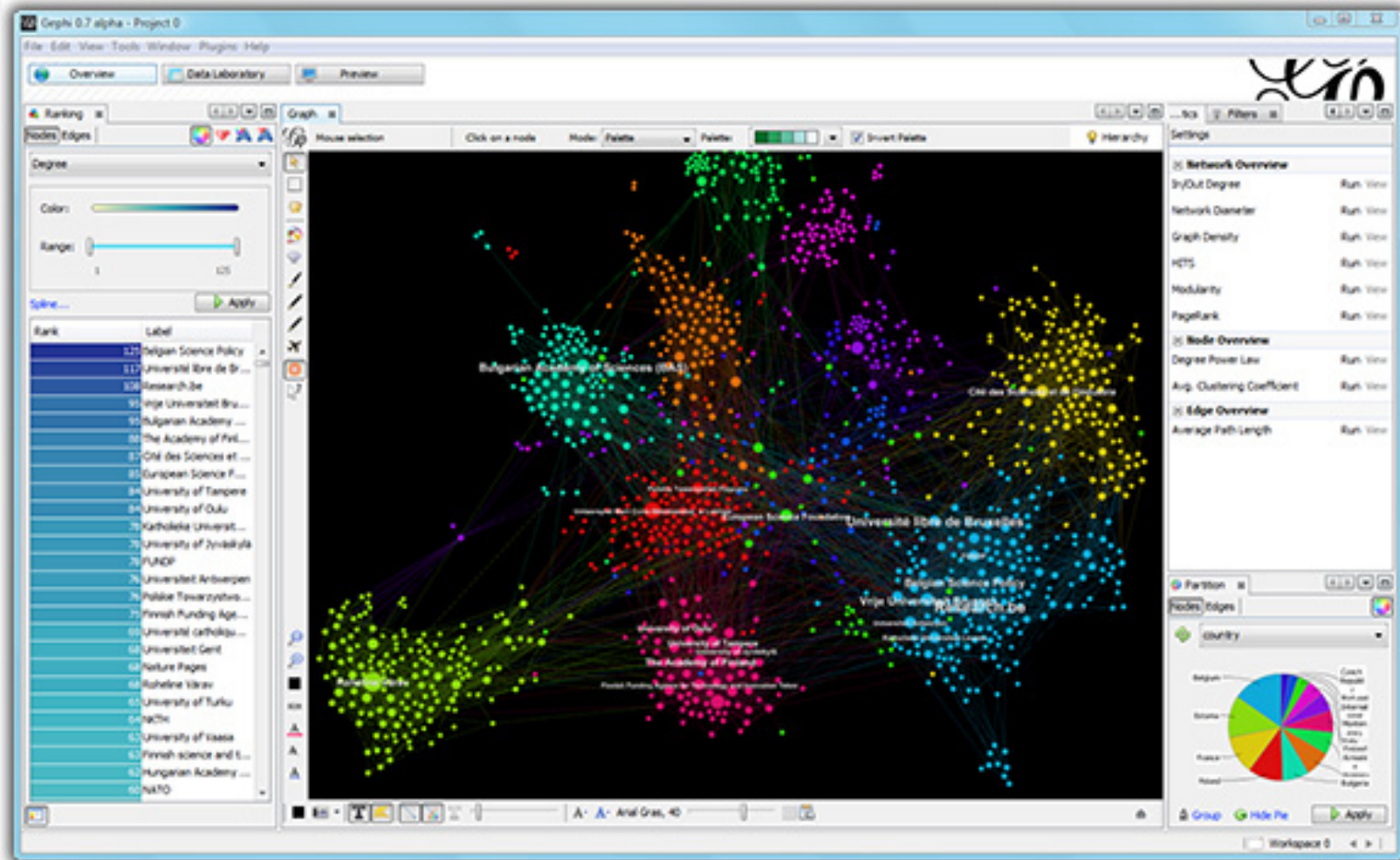
<http://mallet.cs.umass.edu/topics.php>; Alternative: gensim

Stilometrie: stylo for R



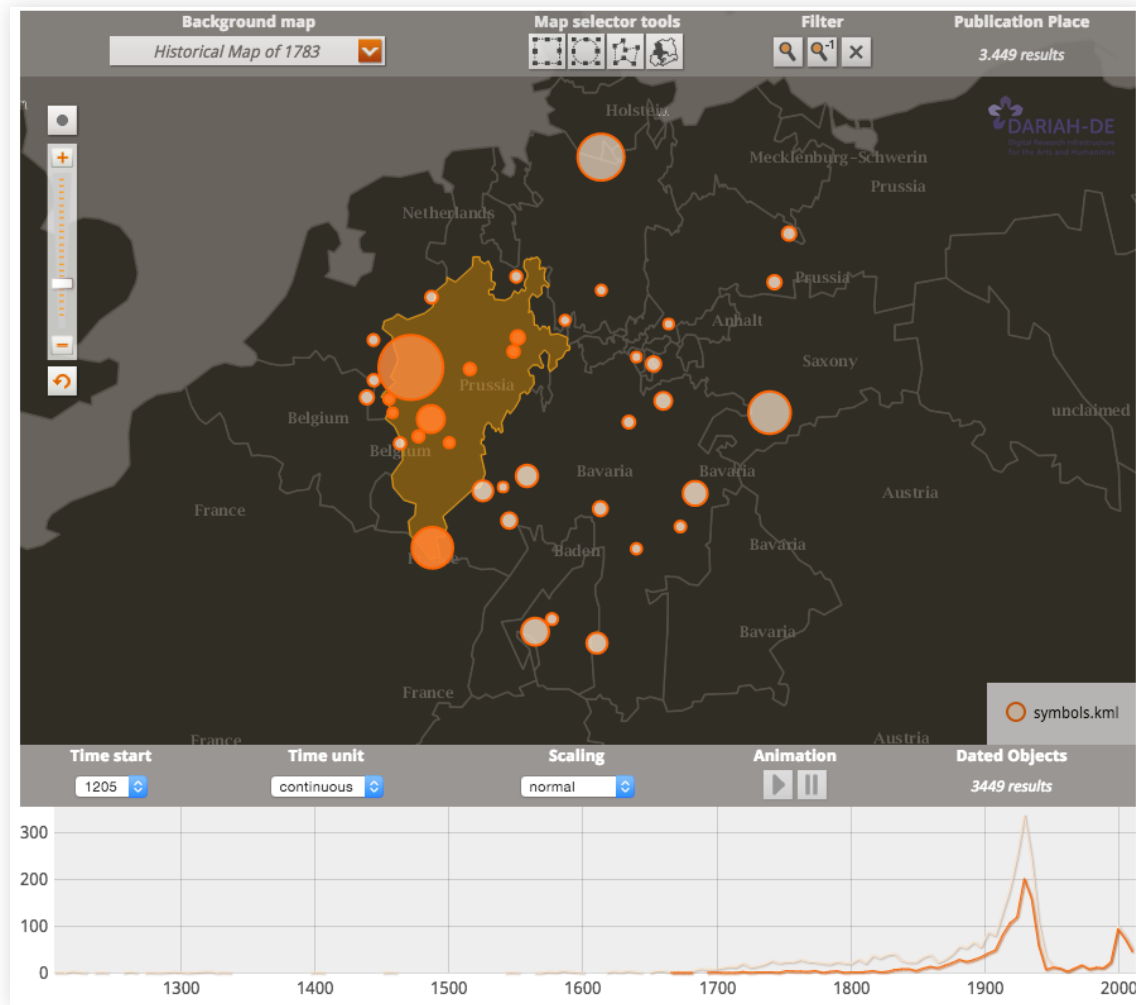
<https://sites.google.com/site/computationalstylistics/>

Netzwerkanalyse: Gephi



<https://gephi.org/>; Alternative: networkX

Kartierung: DARIAH Geobrowser



<https://de.dariah.eu/geobrowser>; Alternative: folium

(3) Stilometrische Autorschaftsattributions

Zwei Typen von ML

5.1

unüberwacht	überwacht
Clustering	Klassifikation
Bilden von Gruppen	Zuordnung zu Klassen
keine Klassen	vorher bekannte Klassen
ein Datensatz	Training/Test/Anwendung
eher explorativ	hypothesengeleitet
Evaluation möglich	Evaluation leicht
Topic Modeling PCA, CA	Annotation OCR, NER

5.2

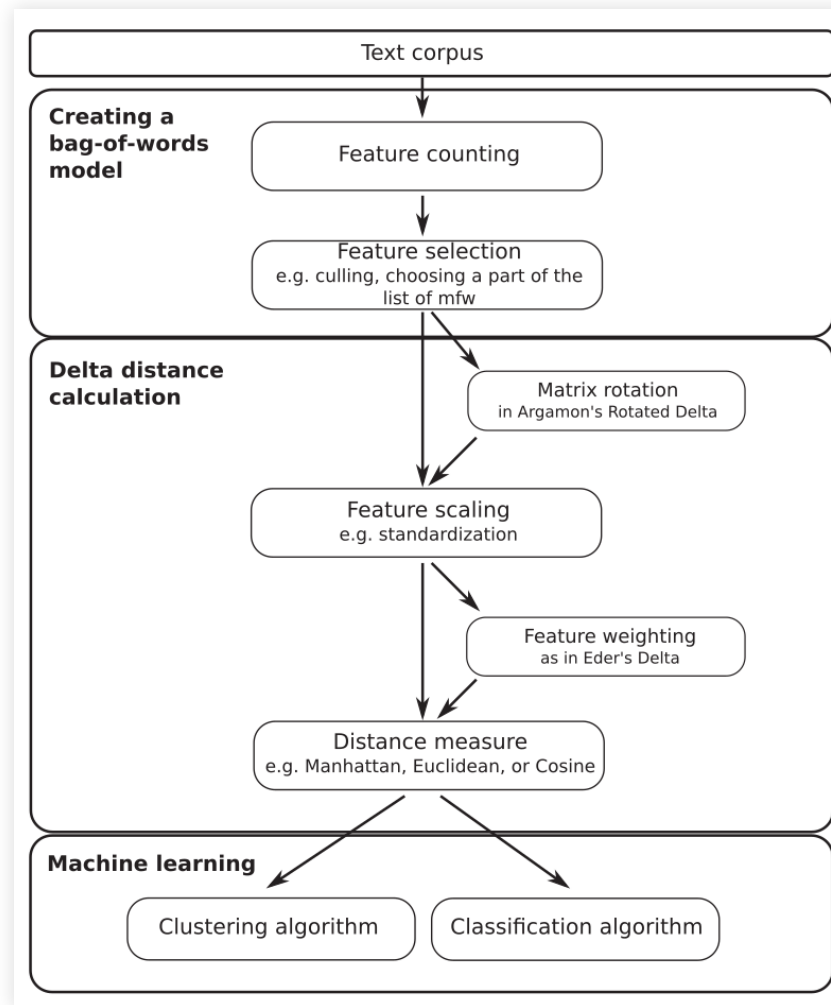
Stilometrie: Definition

Die Stilometrie ist eines von mehreren Verfahren, die dem Bereich der quantitativen Textanalyse zugerechnet werden können. Der Begriff Stilometrie bezeichnet dabei computergestützte Verfahren der Erhebung lexikalischer bzw. stilistischer Merkmale und ihrer Häufigkeiten in Texten, die Nutzung dieser Merkmale und Häufigkeiten als Indikatoren für die mehr oder weniger große Ähnlichkeit von Texten, sowie das Clustering oder die Klassifikation von Texten auf Grundlage dieser Ähnlichkeit.

Anwendungsfälle

- Shakespeare vs. Zeitgenossen
- J.K. Rowling, "The Cuckoo's Calling"
- Elena Ferrante
- Molière vs. Corneille
- uvm.

Stilometrie "step-by-step" (1)



Bildquelle: Steffen Pielström in Evert et al. 2017

Textsammlung (Metadaten)

	A	B	C	D	E	F
1	idno	author	title	year	genre	form
2	tc0189	CorneilleP	Sertorius	1662	Tragédie	vers
3	tc0196	CorneilleP	ConqueteToison	1661	Tragédie	vers
4	tc0200	CorneilleT	Ariane	1672	Tragédie	vers
5	tc0222	CorneilleT	MortAchille	1673	Tragédie	vers
6	tc0226	CorneilleT	Stilicon	1660	Tragédie	vers
7	tc0656	RacineJ	Britannicus	1669	Tragédie	vers
8	tc0661	RacineJ	Phèdre	1677	Tragédie	vers

Drei Autoren: Thomas und Pierre Corneille sowie Racine

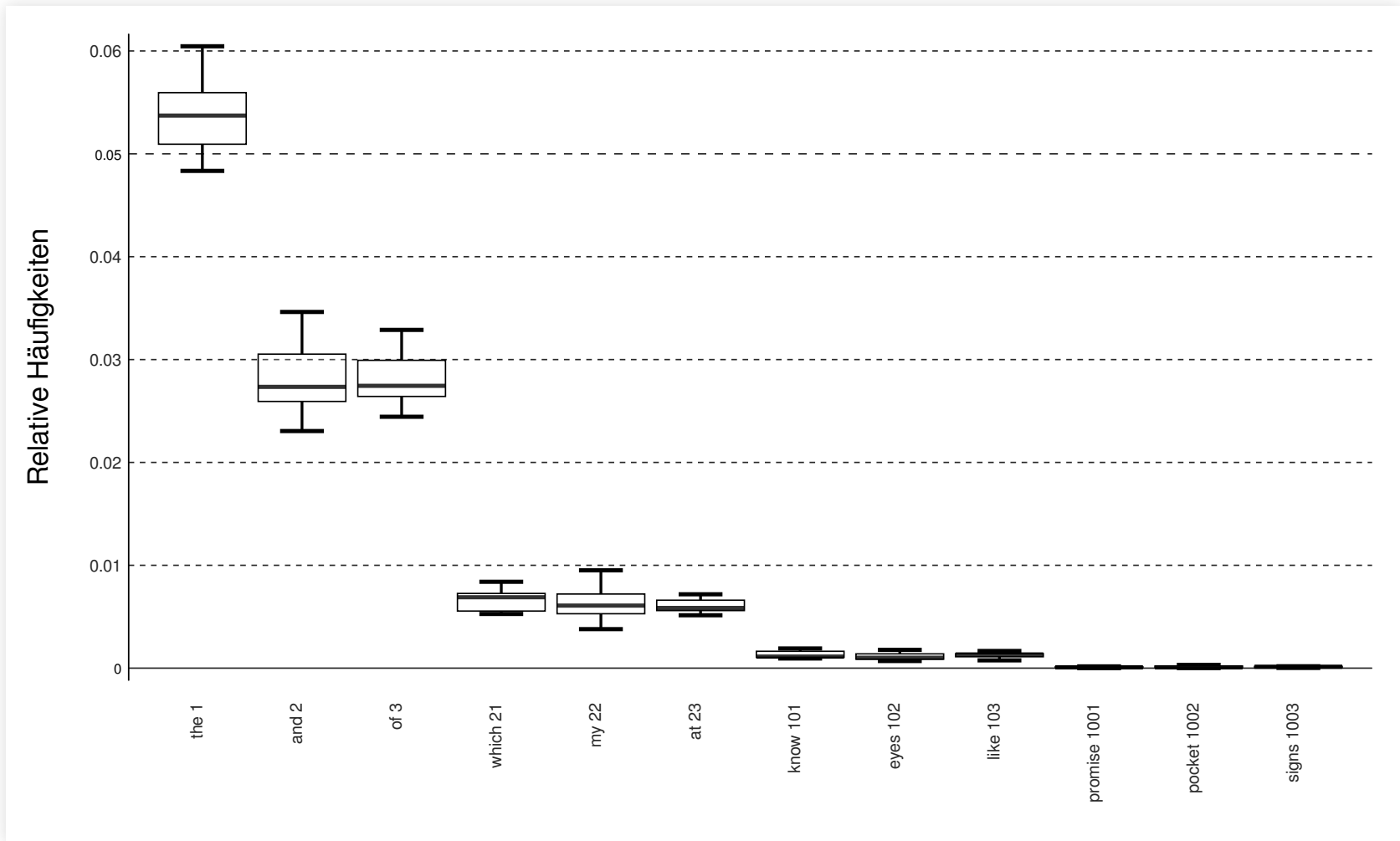
Relative Häufigkeiten

	A	B	C	D	E
1		CorneilleP_tc0189	CorneilleP_tc0196	CorneilleT_tc0222	CorneilleT_tc0226
2	de	3.5666666667	3.7785192533	3.4143340511	2.9700921039
3	et	2.7333333333	2.7640003607	1.9686250384	2.5302701025
4	vous	2.8222222222	2.191360808	1.9686250384	1.3453378868
5	le	1.8777777778	2.0290377852	2.4607812981	2.5613163614
6	à	2.1388888889	2.4844440436	2.159335589	2.2301562662
7	l	1.8888888889	1.5826494725	2.0547523839	2.6803270206
8	que	1.8944444444	1.8351519524	1.956321132	1.5937079582
9	je	1.9611111111	1.3797456939	1.531836358	1.3970816517
10	il	1.4944444444	1.3707277482	1.6425715165	1.7127186174
11	un	1.3888888889	1.4383623411	1.2488465088	1.6247542171
12	la	1.1611111111	1.5105059068	1.4026453399	1.3401635103
13	en	1.5277777778	1.4248354225	1.3903414334	1.5109179344
14	qu	1.4222222222	1.3617098025	1.5379883113	1.5057435579
15	les	1.2611111111	1.6322481739	1.1196554906	0.7709820967
16	d	1.3388888889	1.3662187754	1.1811750231	1.4591741695
17	est	1.1388888889	0.9604112183	1.4887726853	1.2470247335

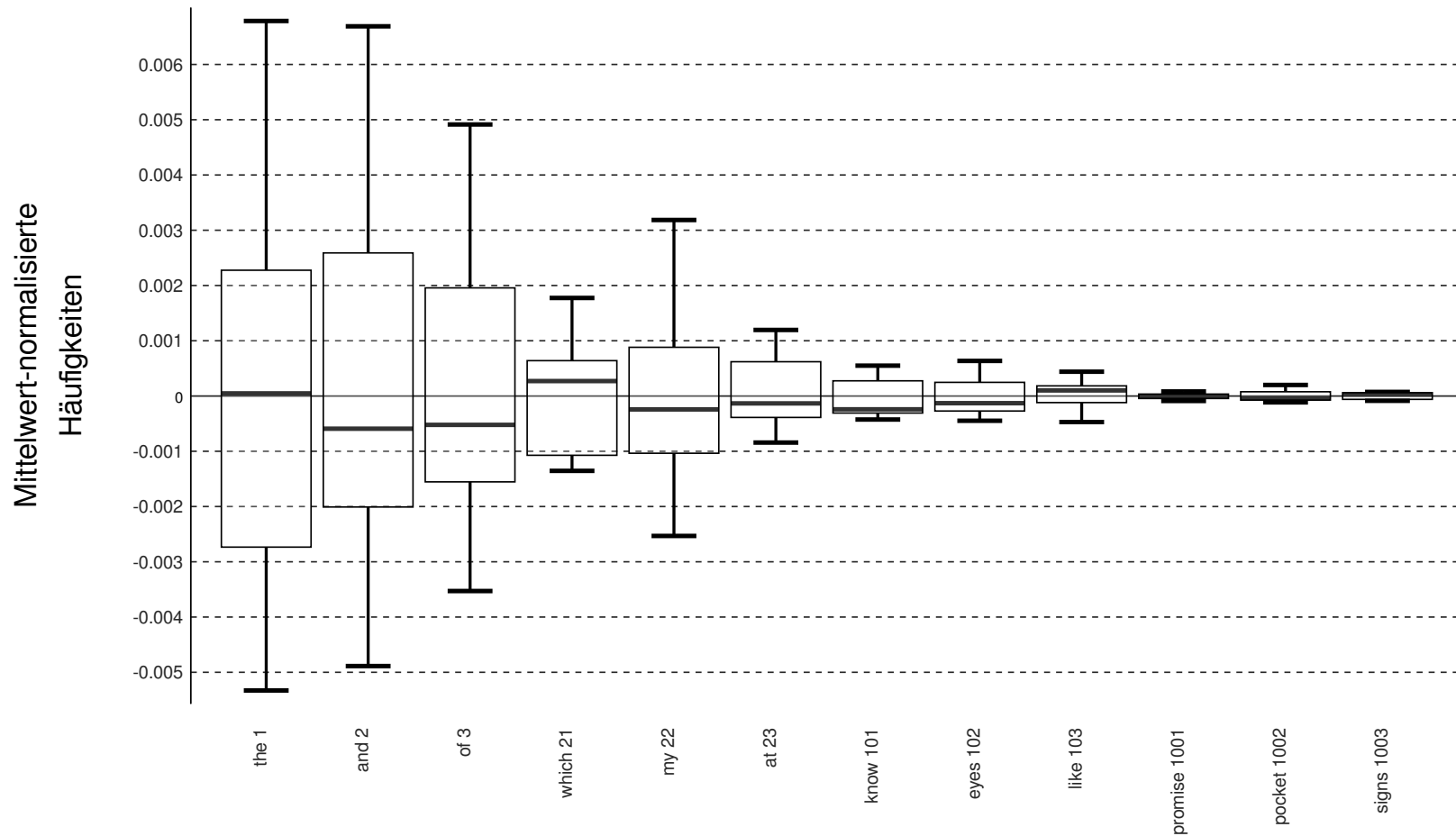
Standardisierung (z-scores)

	A	B	C	D	E
1		CorneilleP_tc0189	CorneilleP_tc0196	CorneilleT_tc0222	CorneilleT_tc0226
2	de	0.1499078729	0.8217586551	-0.3331864738	-1.7420165683
3	et	1.0217971305	1.0964111031	-0.8387654541	0.5277370396
4	vous	1.0822091538	-0.0095321888	-0.3949890303	-1.4736226833
5	le	-1.1206678643	-0.564008828	1.0248706909	1.3948544851
6	à	0.0566725616	1.5330493833	0.1440306397	0.4466104298
7	l	-0.0511188533	-0.7711812699	0.3388768906	1.809793959
8	que	0.7078952579	0.4265714426	1.0014802499	-0.7190026979
9	je	0.6485348916	-1.0497791904	-0.6054842445	-0.9991365133
10	il	0.130171823	-0.3401876567	0.6933373022	0.9600301165
11	un	0.085296816	0.2981974348	-0.5173518384	1.1003032917
12	la	-1.3921359502	1.2926652024	0.463848695	-0.016270901
13	en	0.8709978874	0.4567116762	0.3178923923	0.8031463153
14	qu	0.4674217721	0.2499659708	0.8834356711	0.767561794
15	les	-0.1311125667	0.9118507863	-0.5286288566	-1.508465172
16	d	0.409098182	0.5848113394	-0.6048980286	1.1824532584
17	est	0.0688622182	-0.758505175	1.6908155061	0.5701466176

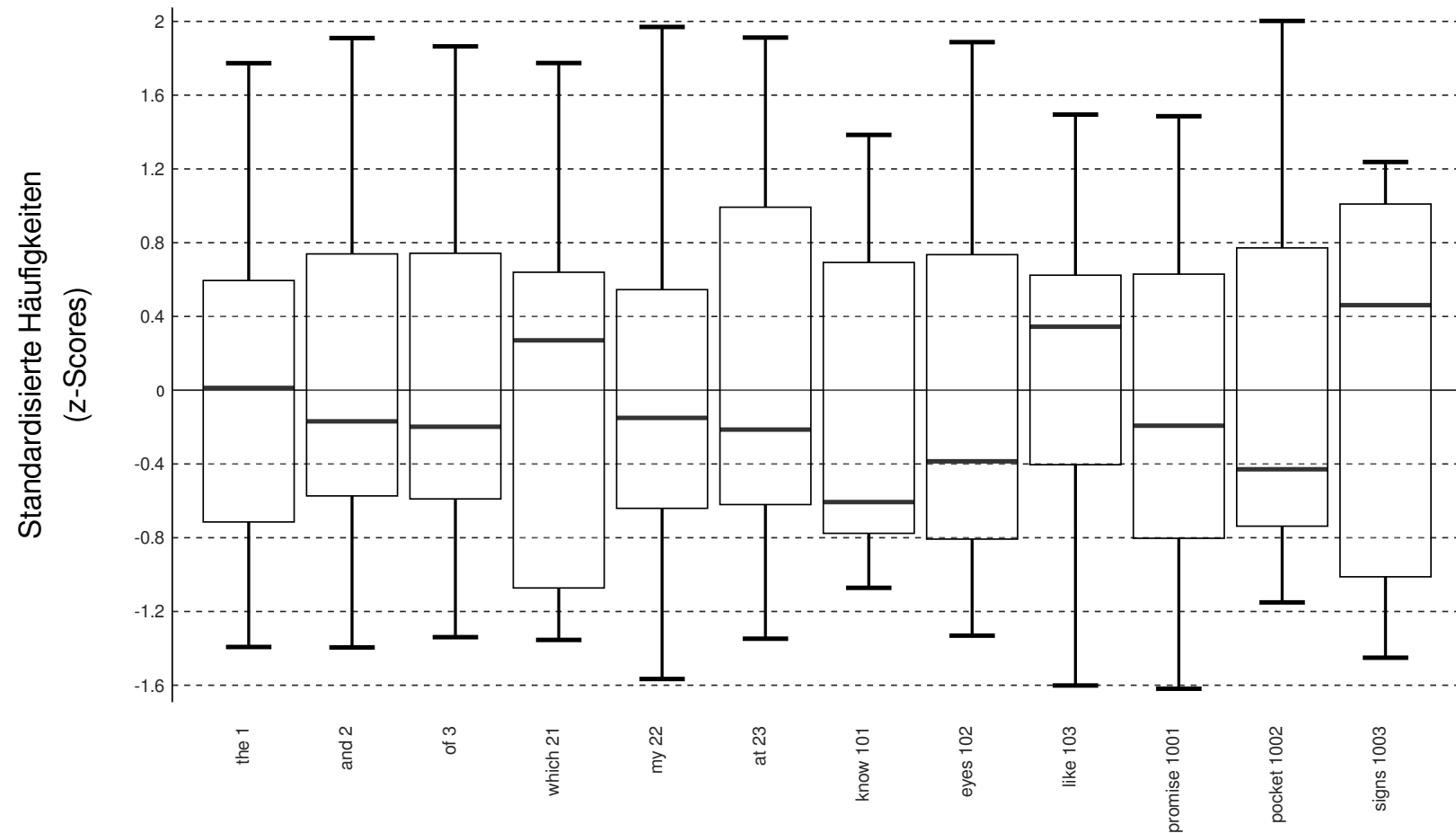
Relative Häufigkeiten



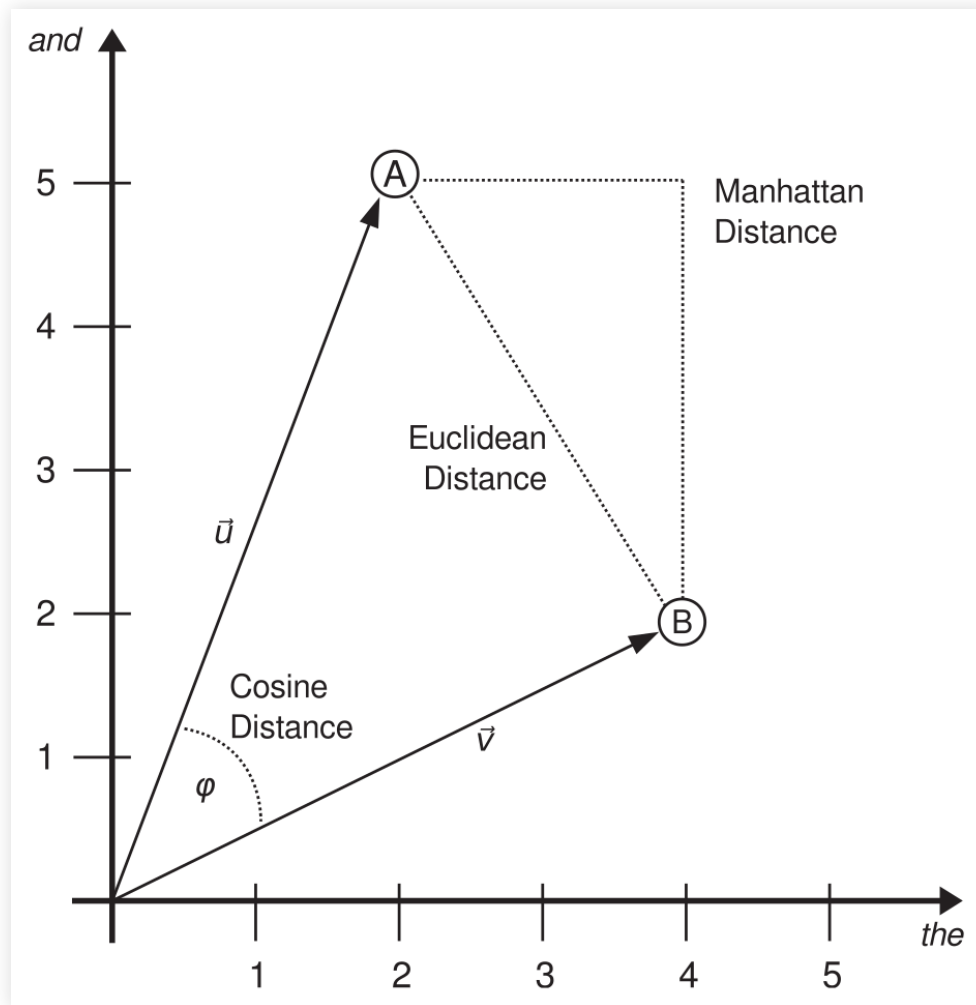
Mittelwert-Normalisierung



Z-Scores (Standardisierung)



Distanzmaße



(Quelle: Digital Humanities: eine Einführung)

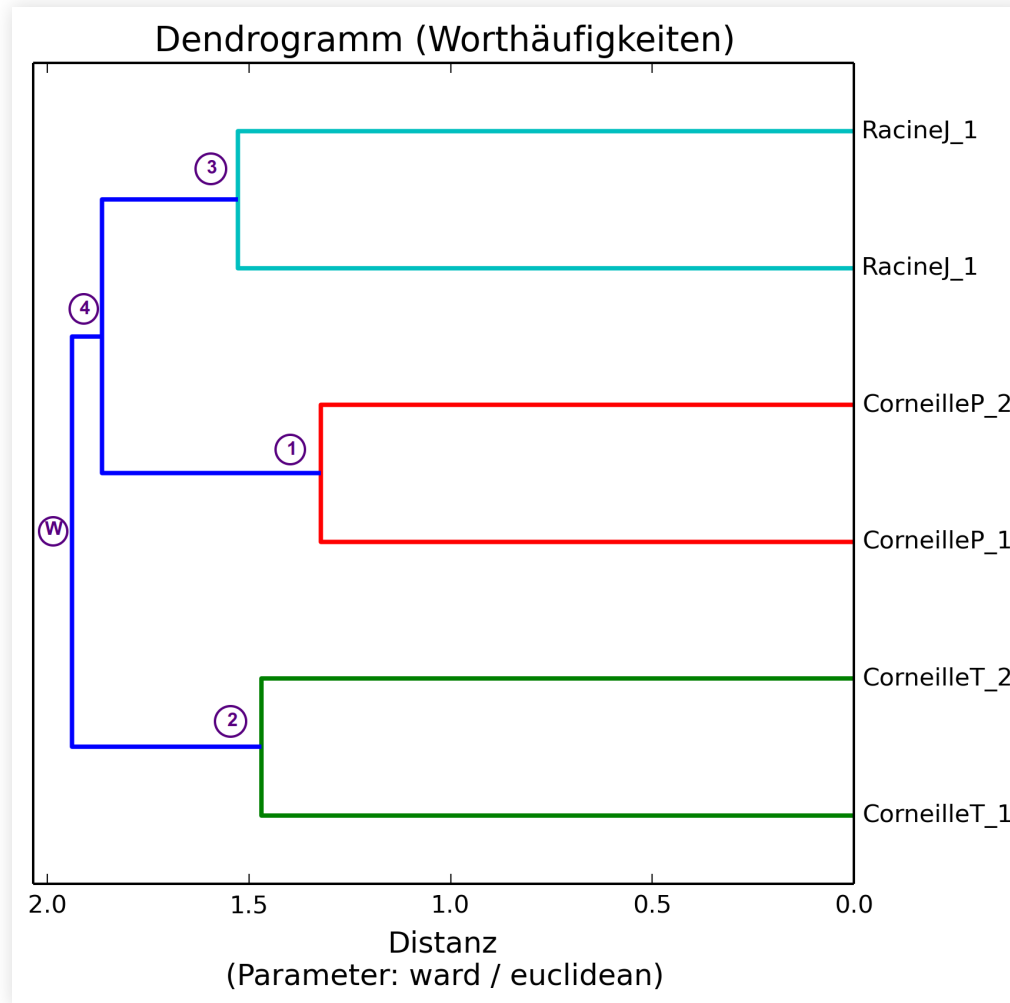
Distanz-Matrix

	A	B	C	D
1		CorneilleP_tc0189	CorneilleP_tc0196	CorneilleT_tc0222
2	CorneilleP_tc0189	0	0.9322543628	1.1478155331
3	CorneilleP_tc0196	0.9322543628	0	1.1417180795
4	CorneilleT_tc0222	1.1478155331	1.1417180795	0
5	CorneilleT_tc0226	1.1997307538	1.1472409053	1.0782741957
6	RacineJ_tc0656	1.1122630299	1.1522653374	1.1985345423
7	RacineJ_tc0661	1.2173503293	1.1504941657	1.1887585769

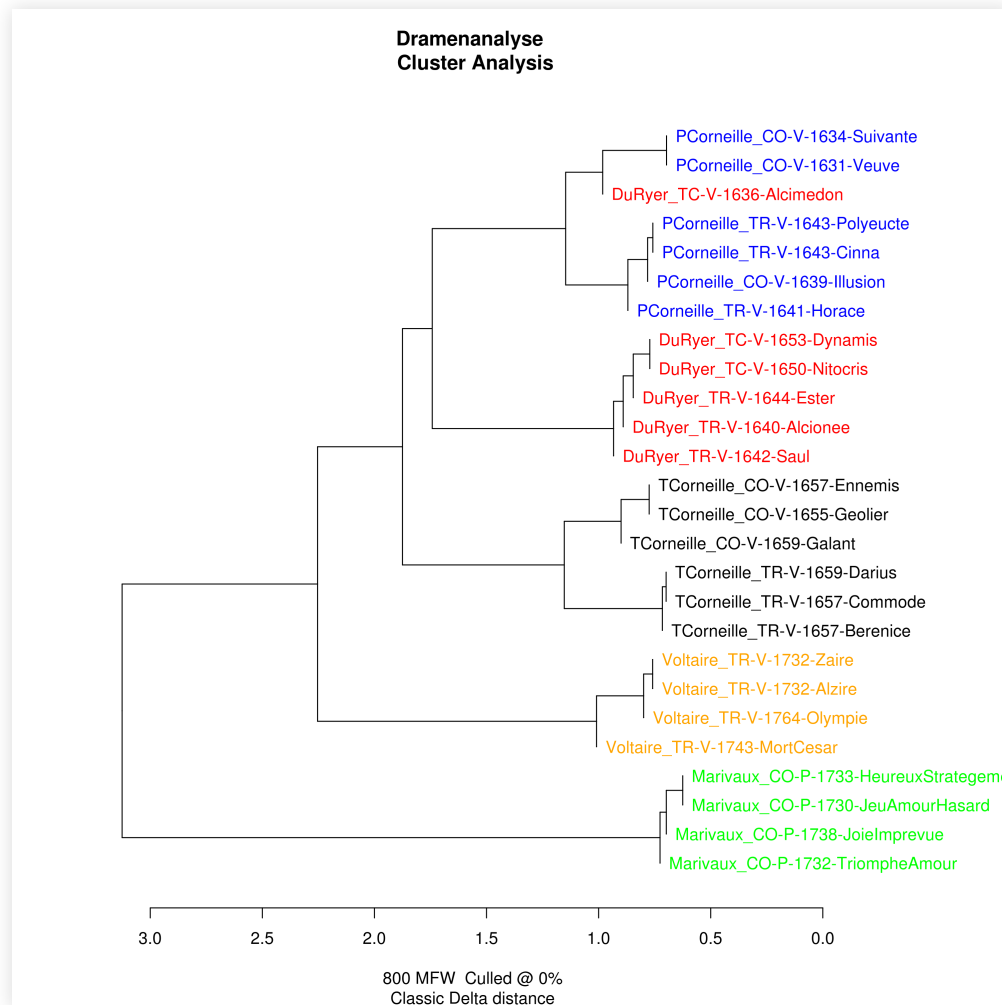
Linkage Matrix

Ast1	Ast2	Distanz	Knoten
CorneilleP_1	CorneilleP_2	1.321	Cluster 1
CorneilleT_1	CorneilleT_2	1.469	Cluster 2
RacineJ_1	RacineJ_2	1.527	Cluster 3
Cluster 1	Cluster 3	1.865	Cluster 4
Cluster 4	Cluster 2	1.938	Wurzel (W)

Dendrogramm



Anwendungsbeispiel



(26 Theaterstücke; 5 Autoren; Vers und Prosa)

Zusammenfassung des Ablaufs

1. Ausgangspunkt: Textsammlung (bspw. in XML-TEI)
2. Vorbereitung: Text extrahieren, Tokenisierung
3. Relativen Häufigkeiten berechnen: => Merkmals-Matrix (Text als Wortvektor)
4. Feature-Auswahl: bspw. Anzahl der häufigsten Wörter
5. Feature-Skalierung (bspw. z-scores): => skalierte Merkmals-Matrix
6. Anwendung eines Distanz-Maßes: => Distanz-Matrix
7. Transformation in eine hierarchische Struktur durch
Cluster Analyse: => Linkage Matrix
8. Visualisierung der Linkage Matrix: => Dendrogramm
9. Interpretation des Dendrogramms: => Aussage

Abschluss

Referenzlektüre

- Christof Schöch, "Quantitative Analyse", in: *Digital Humanities: Eine Einführung*. Hrsg. von Fotis Jannidis, Hubertus Kohle, Malte Rehbein. Stuttgart: Metzler.

Weitere Empfehlungen

- Jannidis, Fotis (2010). "Methoden der computergestützten Textanalyse". *Methoden der literatur- und kulturwissenschaftlichen Textanalyse*, hrsg. von A. Nünning und V. Nünning. Stuttgart & Weimar: Metzler, S. 109–32.
- Alpaydin, E. (2010). *Introduction to Machine Learning*. 2nd ed. Cambridge, Mass: MIT Press.
- Ramsay, Stephen (2011). *Reading Machines: Toward an Algorithmic Criticism*. Urbana Ill.: University of Illinois Press.

Christof Schöch, 2021
<http://www.christof-schoech.de>

Lizenz: Creative Commons Attribution 4.0