FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

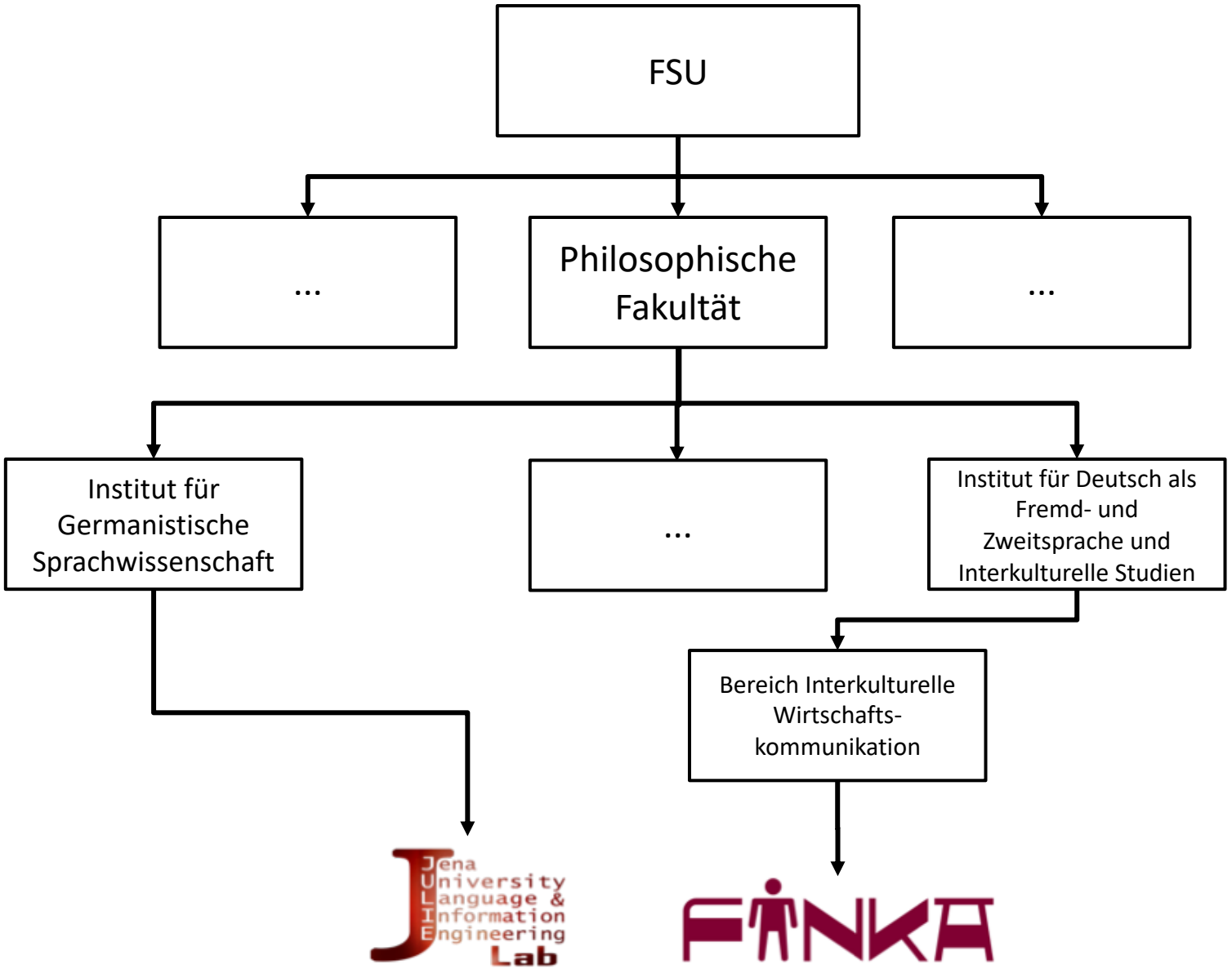
Jena University Language & Information Engineering Lab

Quantifizierung von Emotionen in Historischer Sprache

Anwendungen und Methodische Grundlagen

Sven Büchel

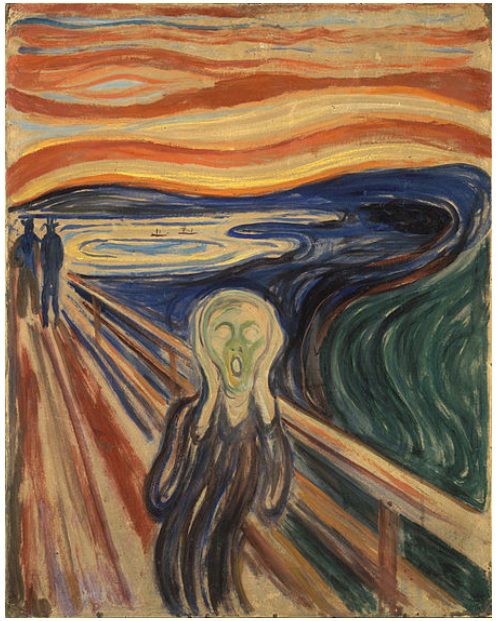
Jena University Language & Information Engineering
(JULIE) Lab
http://www.julielab.de
Friedrich-Schiller-Universität Jena


FSU
...
Philosophische Fakultät
...
Institut für Germanistische Sprachwissenschaft
...
Institut für Deutsch als Fremd- und Zweitsprache und Interkulturelle Studien
Bereich Interkulturelle Wirtschafts-kommunikation
Jena University Language & Information Engineering Lab
FINKA

- Studium der Physik, Germanistik und Soziologie in Jena
- B.A. erlangt 2016 mit „Automatische Analyse von Emotionen in Geschäfts- und Nachhaltigkeitsberichten“
- Ko-Betreut vom Lehrstuhl für ABWL / Organisation, Führung und HRM (Prof. Dr. Peter Walgenbach)
- Ab Sommer 2016 Direktpromotion am JULIE Lab (Prof. Dr. Udo Hahn)
- Forschungsschwerpunkt: Methodenentwicklung zur Automatischen Messung von Emotionen in Text

- Studium der Physik, Germanistik und Soziologie in Jena
- B.A. erlangt 2016 mit „Automatische Analyse von Emotionen in Geschäfts- und Nachhaltigkeitsberichten“
- Ko-Betreut vom Lehrstuhl für ABWL / Organisation, Führung und HRM (Prof. Dr. Peter Walgenbach)
- Ab Sommer 2016 Direktpromotion am JULIE Lab (Prof. Dr. Udo Hahn)
- Forschungsschwerpunkt: Methodenentwicklung zur Automatischen Messung von Emotionen in Text
- Zusammen mit Johannes Hellrich: Emotionen in historischer Sprache

Geisteswissenschaften und Emotionen

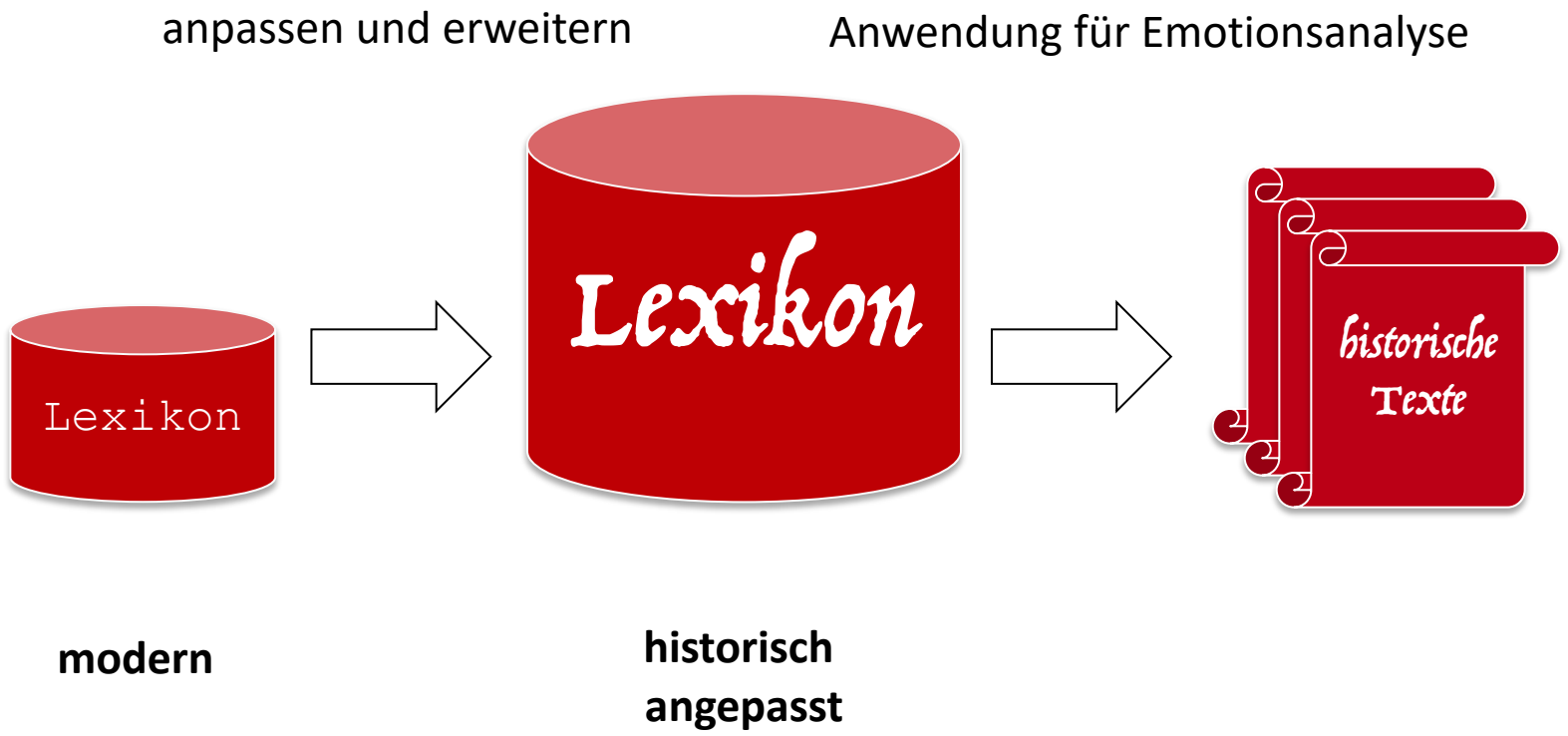


https://de.wikipedia.org/wiki/Walther_von_der_Vogelweide
https://de.wikipedia.org/wiki/Laokoon-Gruppe
https://de.wikipedia.org/wiki/Der_Schrei
http://www.br.de/telekolleg/faecher/deutsch/literatur/goethe-weimarer-klassik-100.html

Unterschiedliche Datendichten

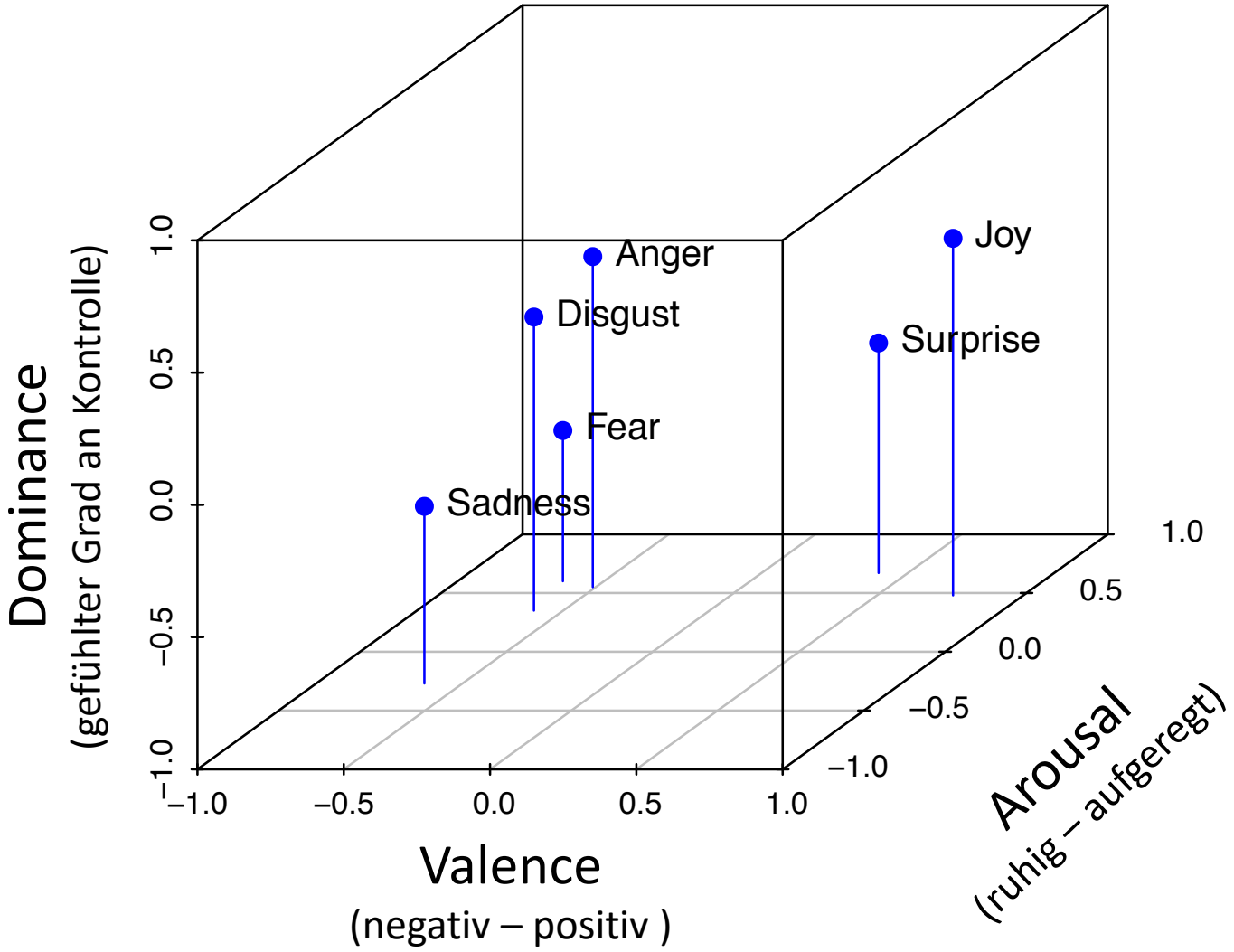
Zeitgenössische Sprachdaten	Historische Sprachdaten
	

Methodisches Vorgehen



Methoden

Formale Darstellung von Emotionen: VAD



Emotionslexika

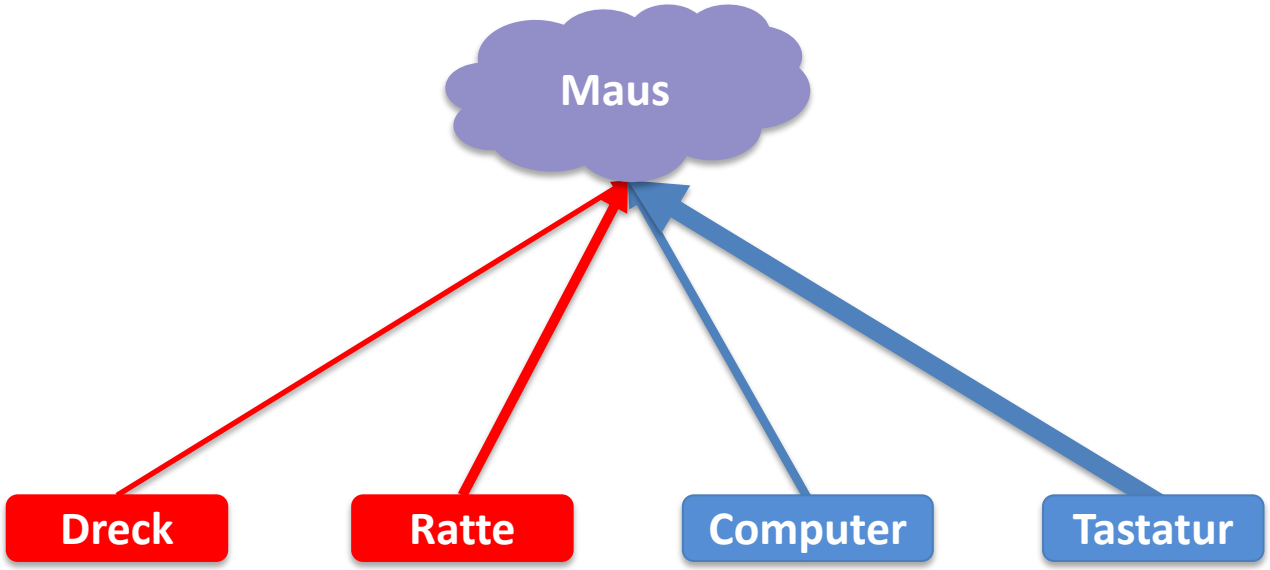
- Lexikalische Datensätze
- Kodiert Emotionen einzelner Wörter
- Kontextunabhängig
- Zahlreiche Datensätze in Psychologie aufgebaut

Lemma	Valence	Arousal	Dominance
Sonnenschein	8.1	5.3	5.4
Terrorismus	1.6	7.4	2.7
Ekstase	6.9	7.8	3.0

(9-Punkt-Skala)

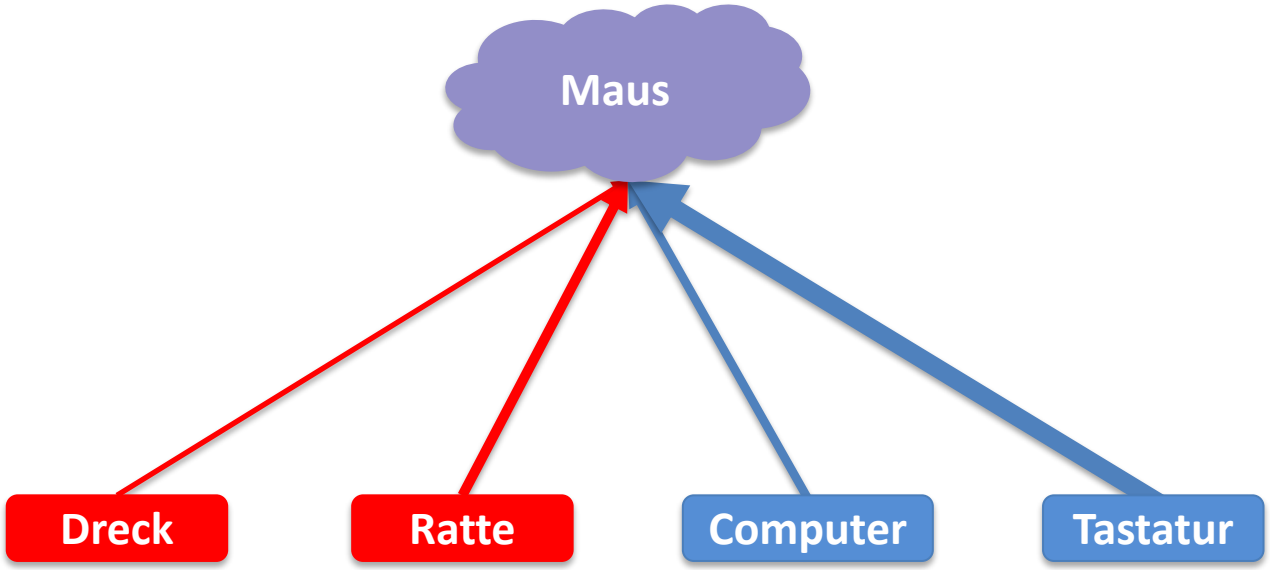
Automatische Erweiterung von Emotionslexika

- Basierend auf sog. *seed*-Wörtern
- Turney-Littman-Algorithmus



Automatische Erweiterung von Emotionslexika

- Basierend auf sog. *seed*-Wörtern
- Turney-Littman-Algorithmus



- Aber: Wie bestimmen wir Wort-Ähnlichkeiten?
  → **Worteinbettungen**

Wortsemantik und Wortumfeld

„Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache.“

Wittgenstein, Philosophische Untersuchungen, 1953

“You shall know a word by the company it keeps!”

Firth, A synopsis of Linguistic Theory, 1957

Wortsemantik und Wortumfeld

Er liest ein Gedicht.
Susanne liest einen Roman.
Der Roman hat 100 Seiten.
Ihr Gedicht hat 3 Seiten.
Susanne hört eine Oper.
Peter hört ein Lied.
Das Lied ist in d-Moll.
Die Oper ist in d-Moll.

Wortsemantik und Wortumfeld

Er liest ein Gedicht.
Susanne liest einen Roman.
Der Roman hat 100 Seiten.
Ihr Gedicht hat 3 Seiten.
Susanne hört eine Oper.
Peter hört ein Lied.
Das Lied ist in d-Moll.
Die Oper ist in d-Moll.

Wortsemantik und Wortumfeld

Er liest ein Gedicht.
Susanne liest einen Roman.
Der Roman hat 100 Seiten.
Ihr Gedicht hat 3 Seiten.
Susanne hört eine Oper.
Peter hört ein Lied.
Das Lied ist in d-Moll.
Die Oper ist in d-Moll.

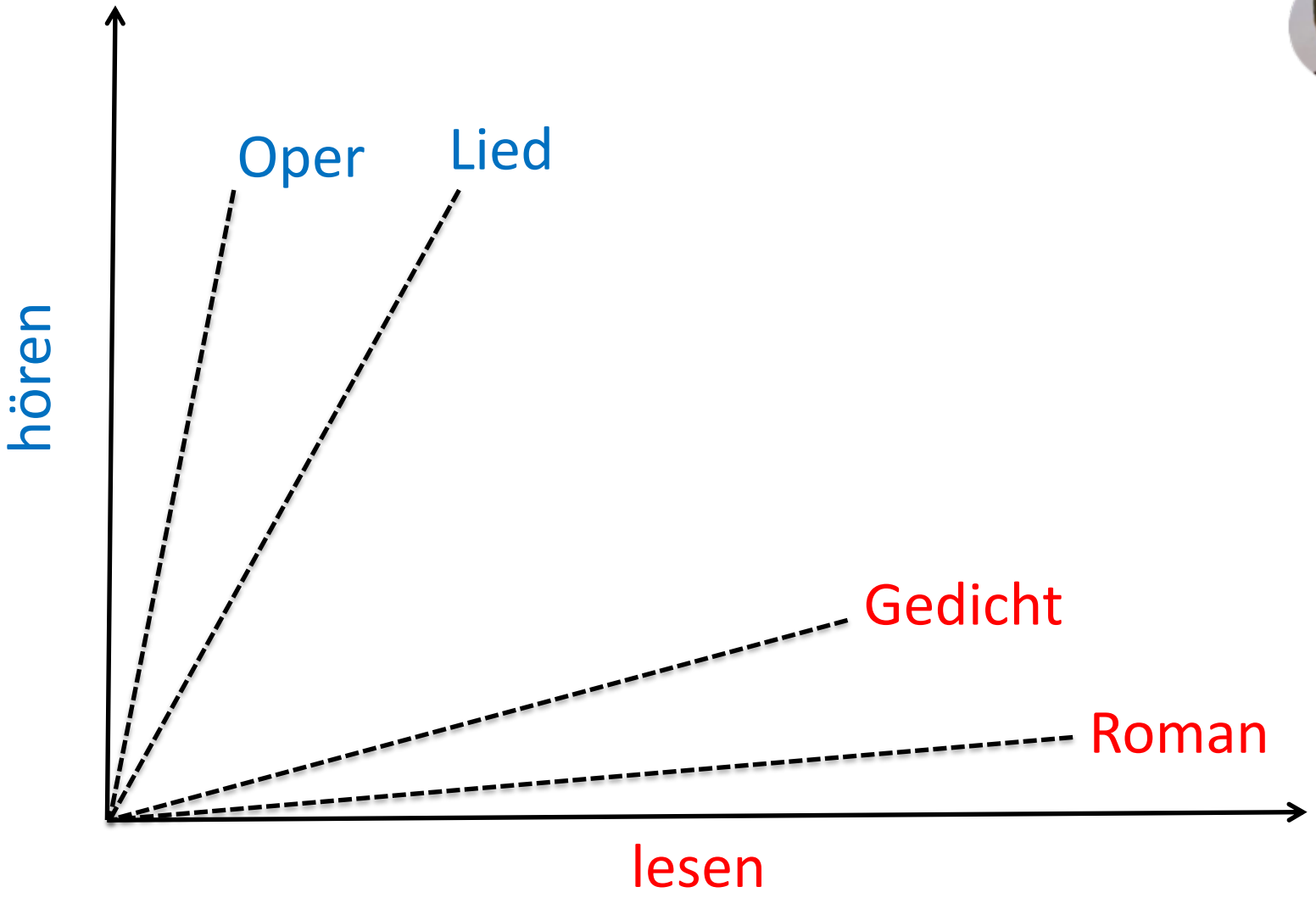
Kookkurrenz-Tabelle

	lesen	Seiten	kaufen	essen	hören	
Roman	98	60	3	0	2	
Gedicht	67	10	1	0	8	…
Oper	4	8	0	0	38	
Lied	12	1	2	0	47	
⋮						

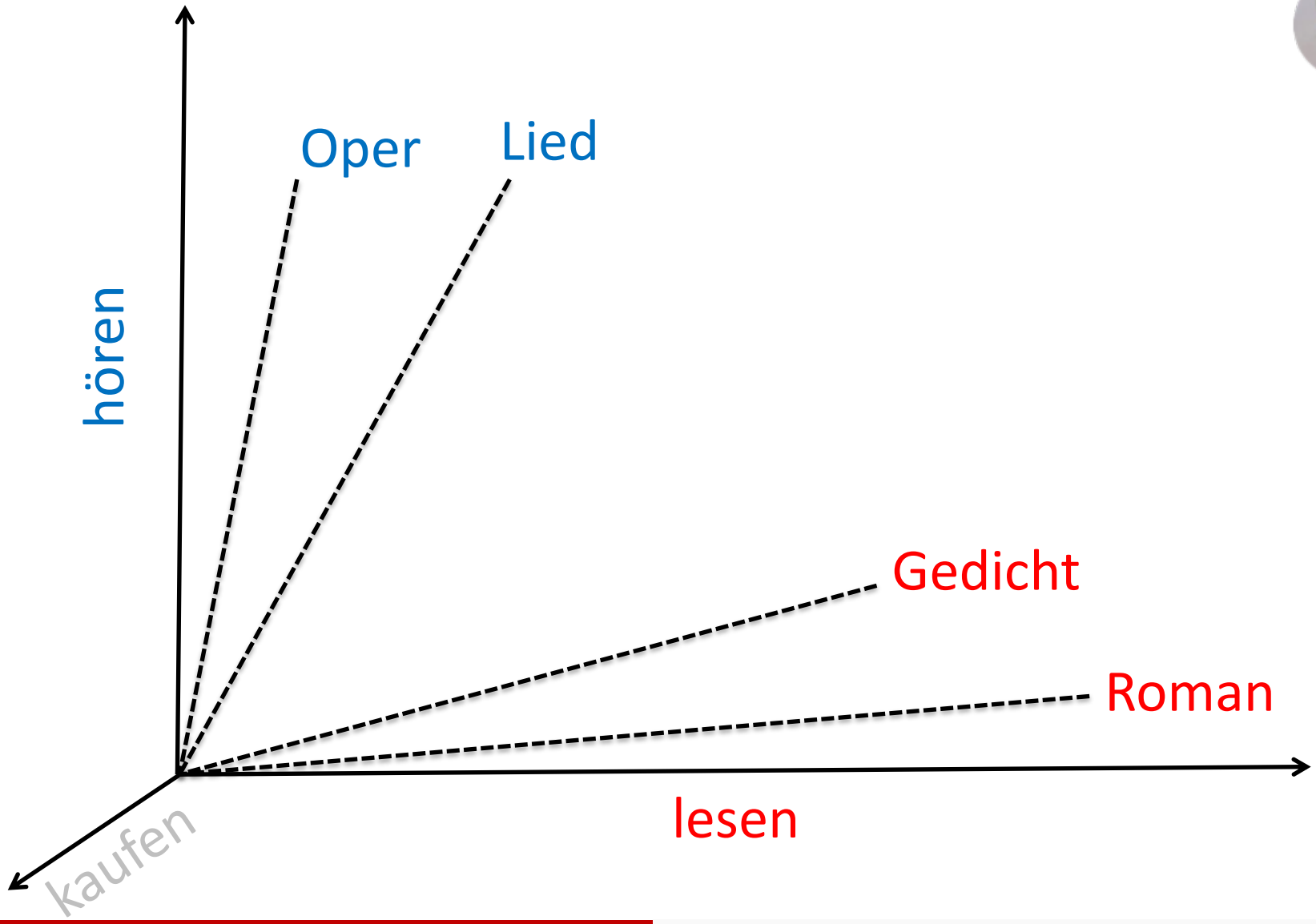
Wortvektoren

	lesen	Seiten	kaufen	essen	hören	
Roman	98	60	3	0	2	
Gedicht	67	10	1	0	8	…
Oper	4	8	0	0	38	
Lied	12	1	2	0	47	
⋮						

Wörter im Vektorraum



Wörter im Vektorraum



Dimensionsreduktion

	lesen	Seiten	kaufen	essen	hören	
Roman	98	60	3	0	2	
Gedicht	67	10	1	0	8	…
Oper	4	8	0	0	38	
Lied	12	1	2	0	47	
⋮						

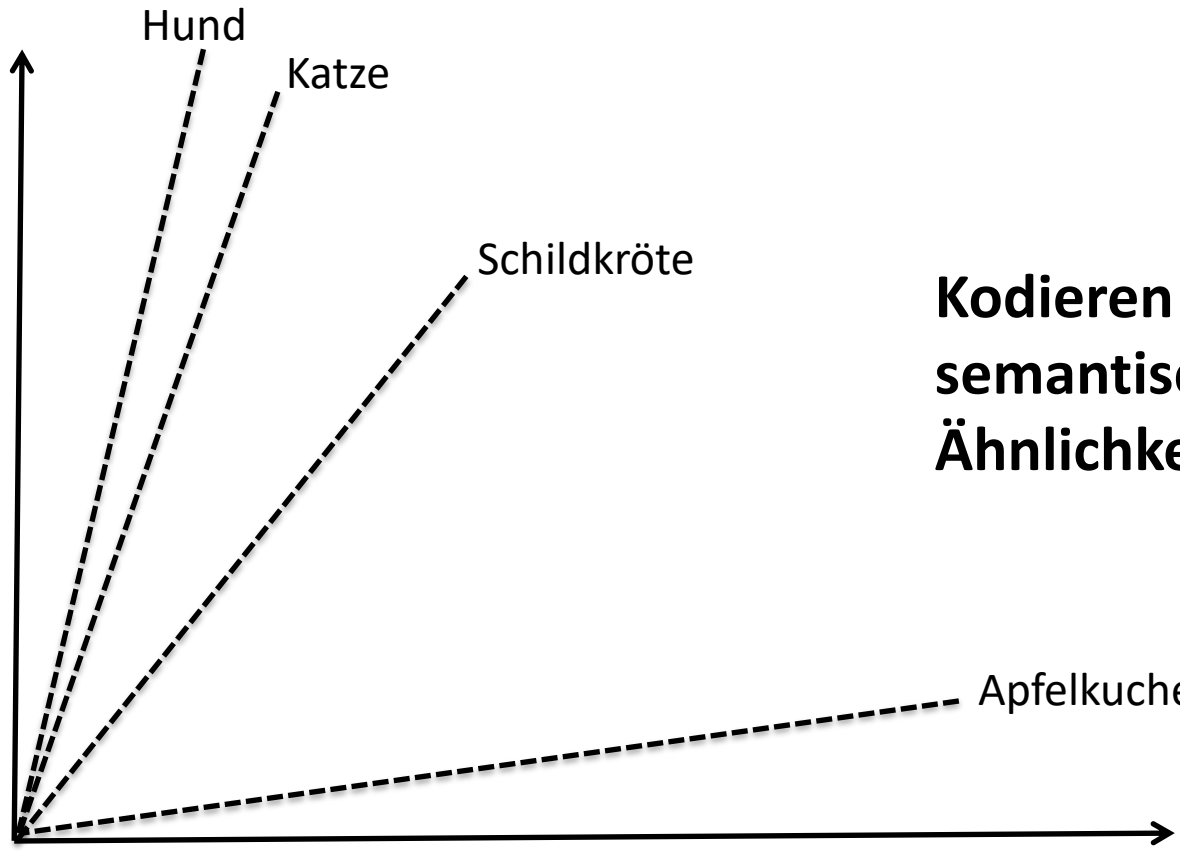
- In der Praxis 100.000 x 100.000 = 10 Milliarden Zellen
- Viele Spalten wenig informativ (z.B. Synonymie)

Dimensionsreduktion

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	
Roman	2.37	-0.334	0.385	-2.84	0.015	
Gedicht	1.576	2.002	-1.23	-1.19	0.385	...
Oper	-0.778	0.624	-2.846	3.125	-2.967	
Lied	1.113	-4.12	0.112	-1.235	-1.285	
⋮						

- Reduktion auf die 300 relevantesten Dimensionen
- Nicht mehr direkt interpretierbar
- Verschiedene mathematische Verfahren (z.B. Neuronale Netze) → **Worteinbettungen**

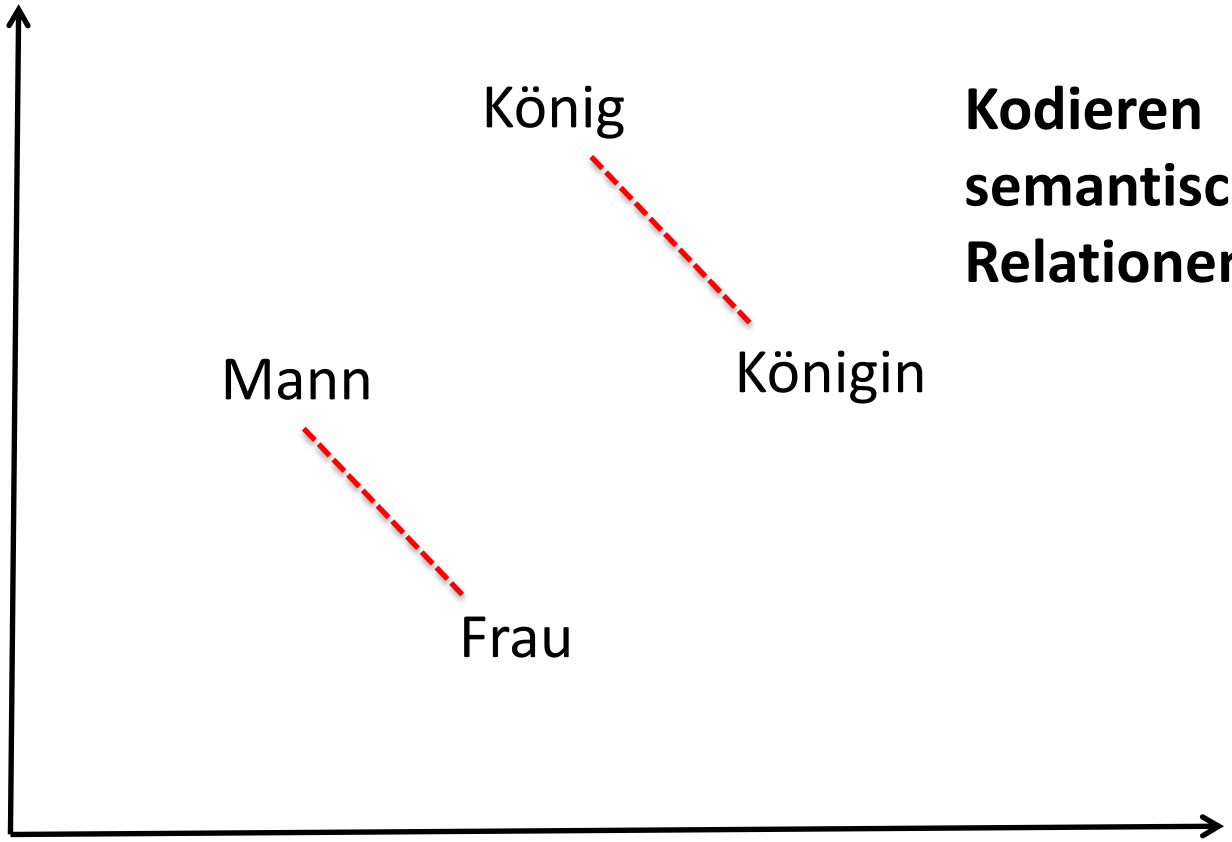
Worteinbettungen



Kodieren semantische Ähnlichkeit!

cos(Hund, Katze) > cos(Hund,Schildkröte) > cos(Hund, Apfelkuchen)

Worteinbettungen

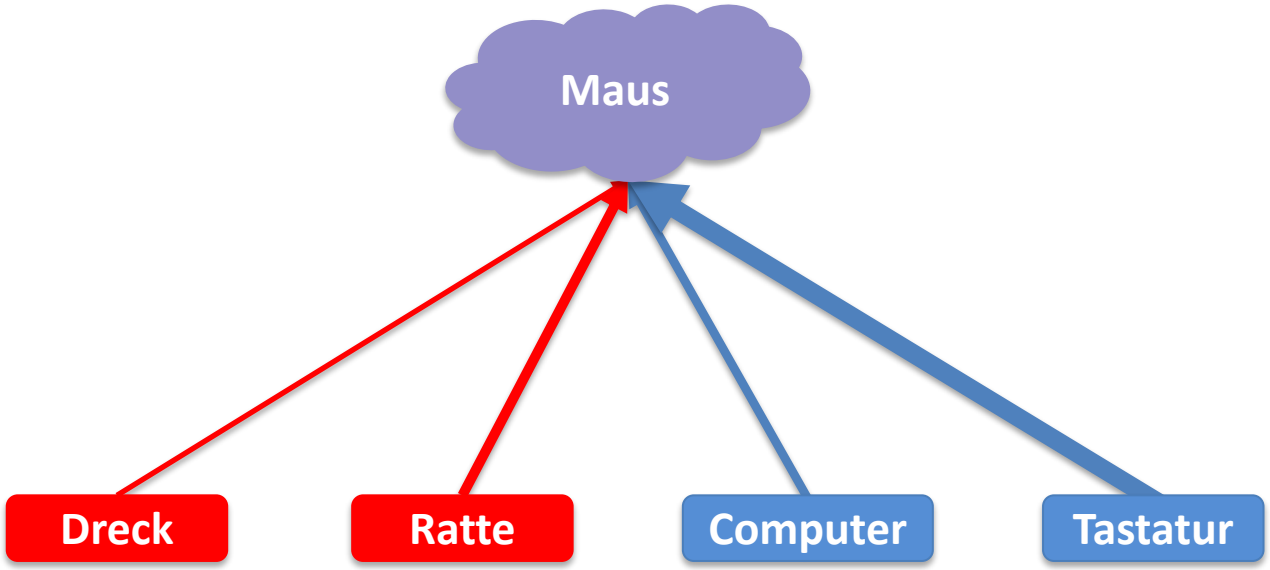


Kodieren semantische Relationen!

Mann – Frau + König = Königin

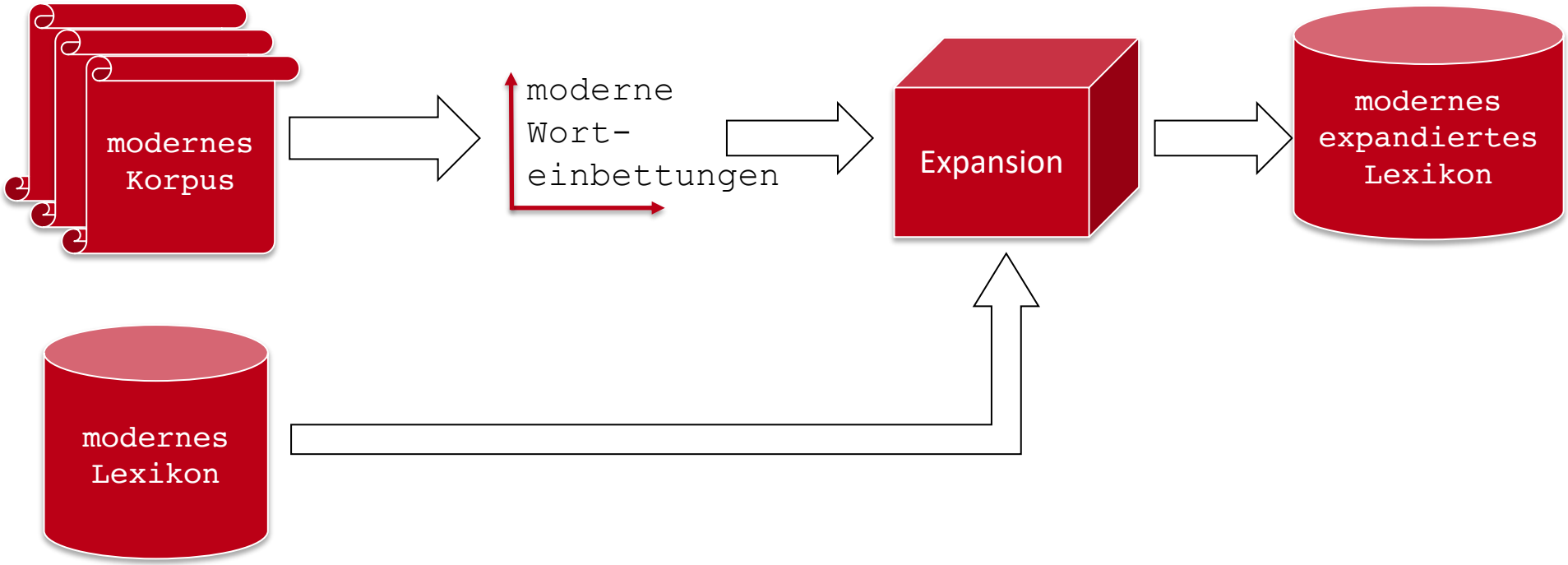
Automatische Erweiterung von Emotionslexika

- Basierend auf sog. *seed*-Wörtern
- Turney-Littman-Algorithmus



- Wortähnlichkeit durch Worteinbettungen!

Synchrone Lexikon-Expansion



Diachrone Lexikon-Expansion

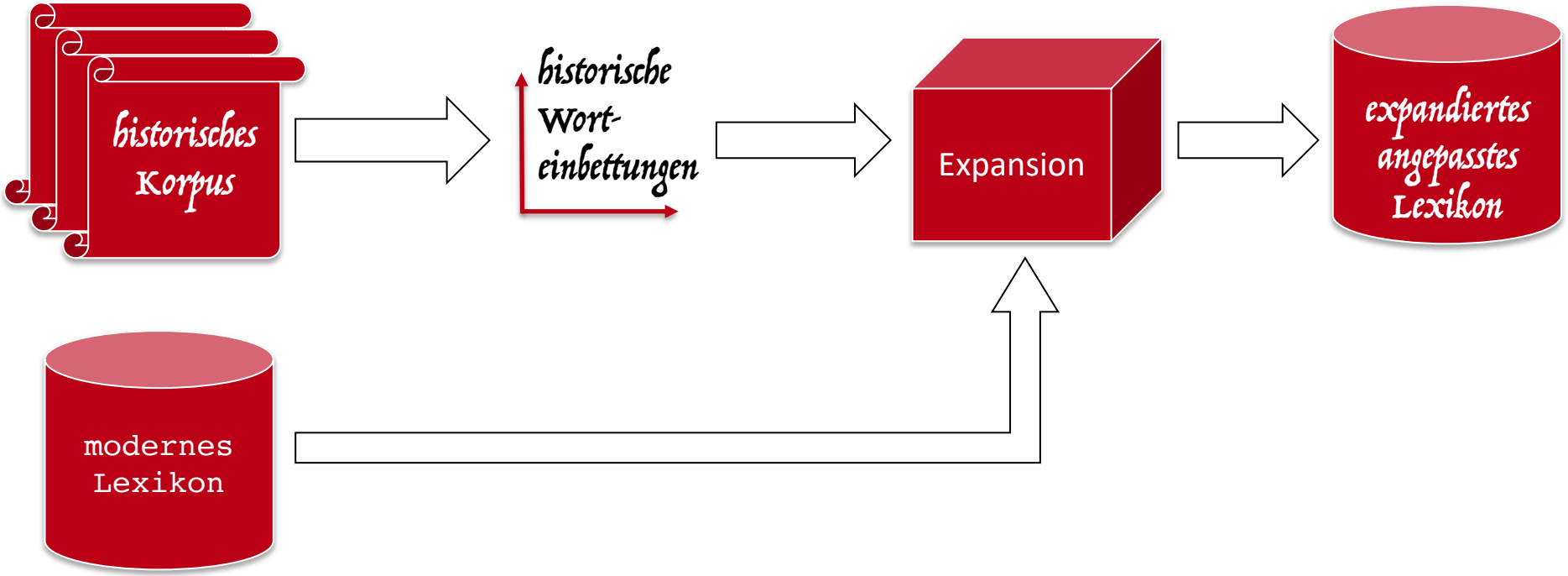


Illustration des Anpassungsschritts (vorher)

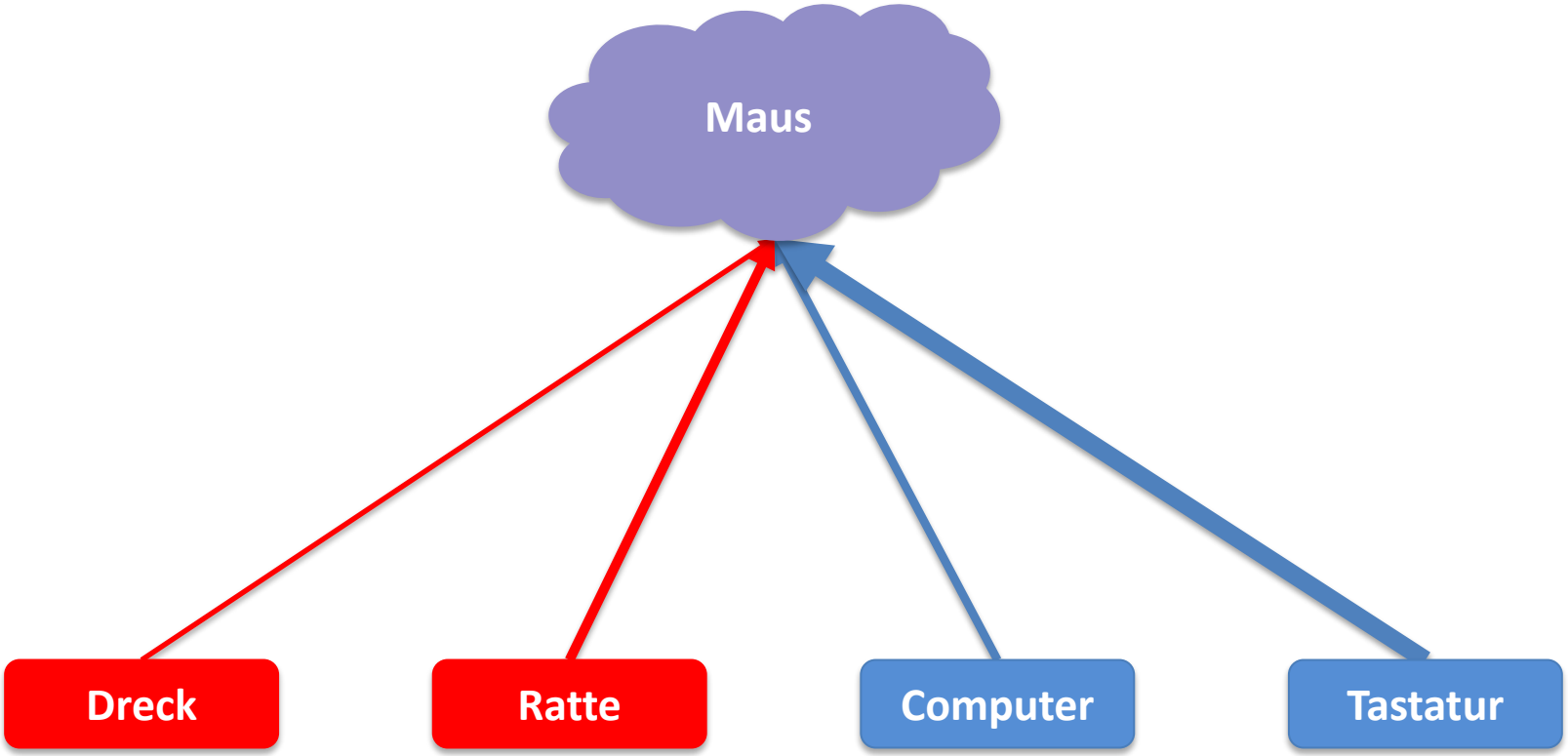
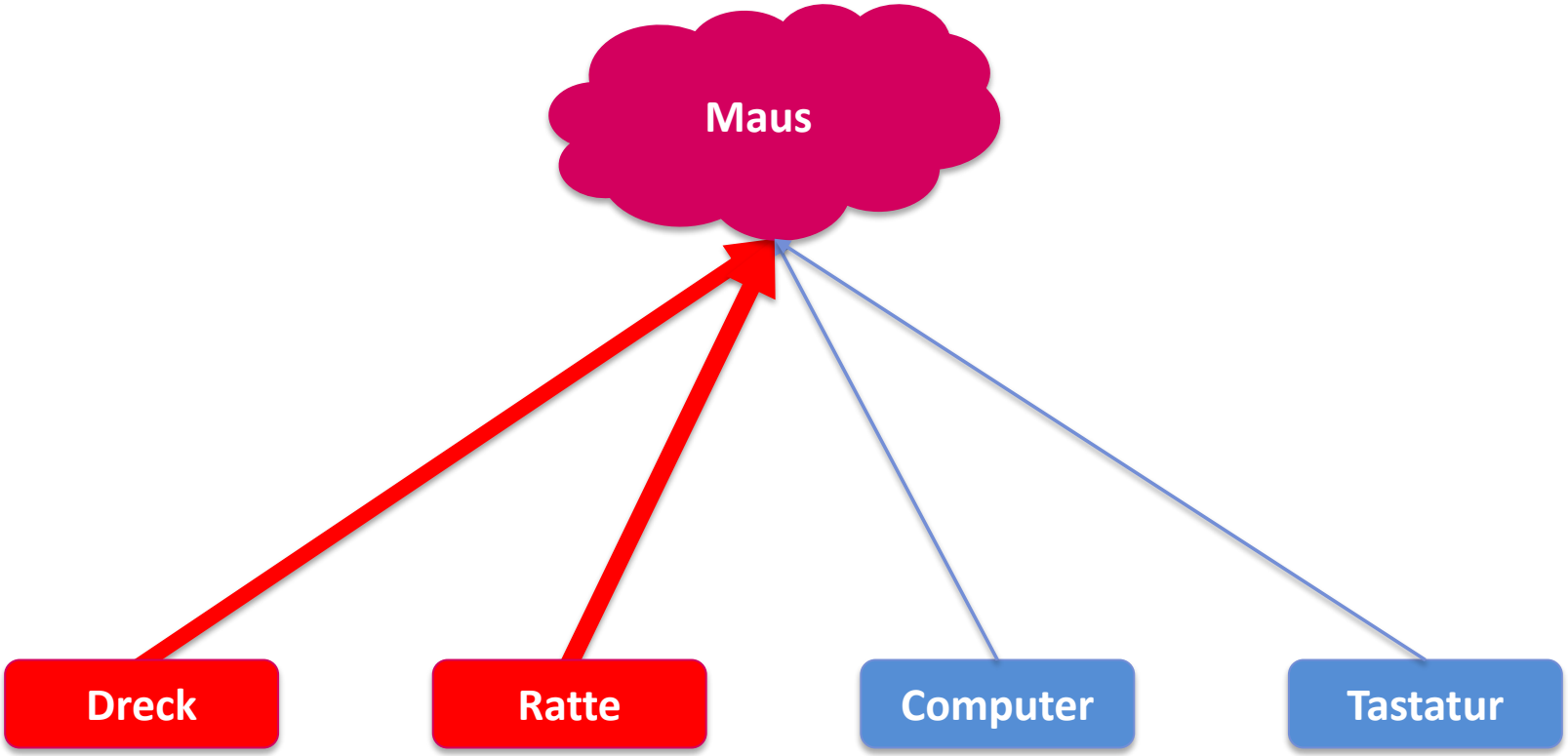
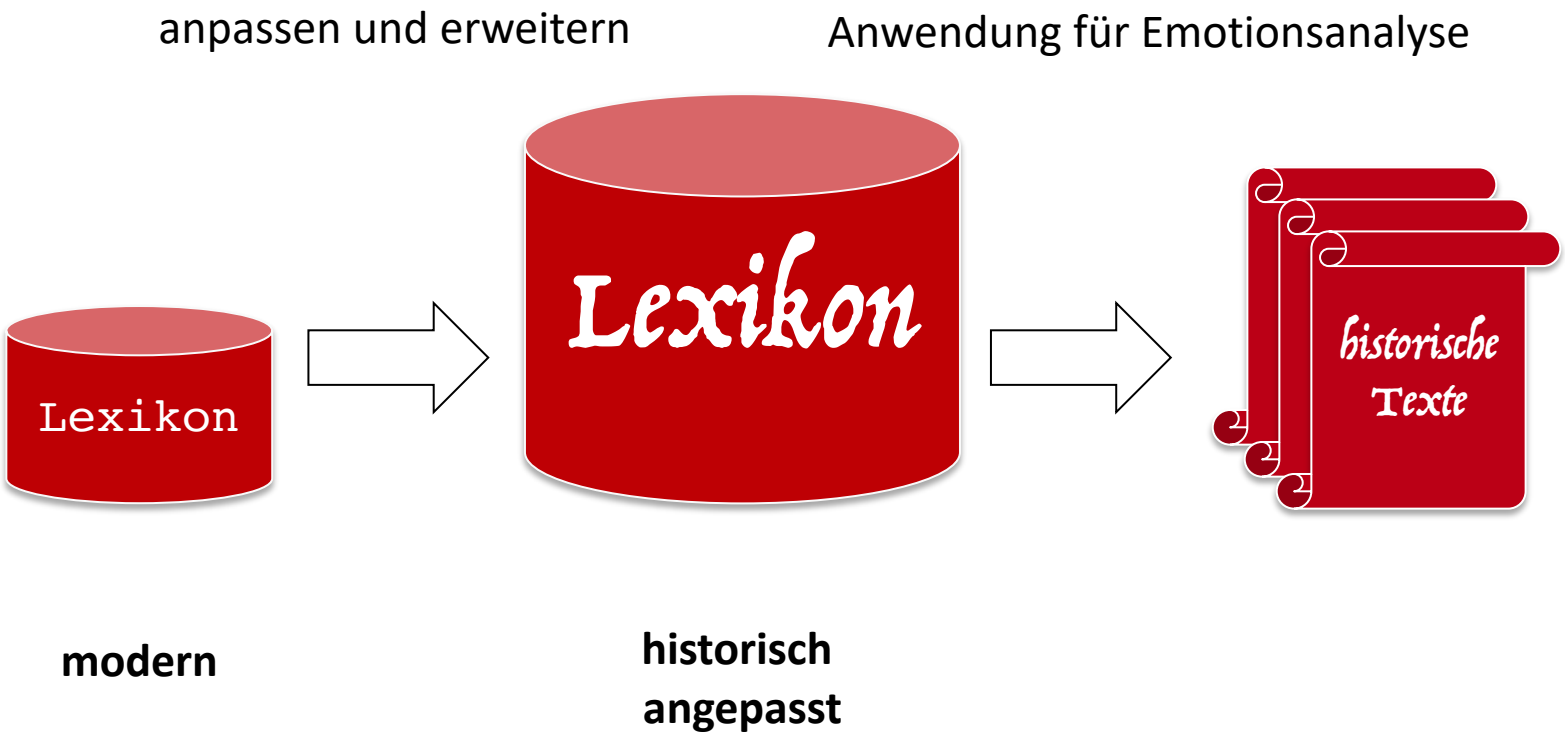


Illustration des Anpassungsschritts (nachher)



Experiment I
(Quantifizierung historischer Wortemotionen)

Methodisches Vorgehen

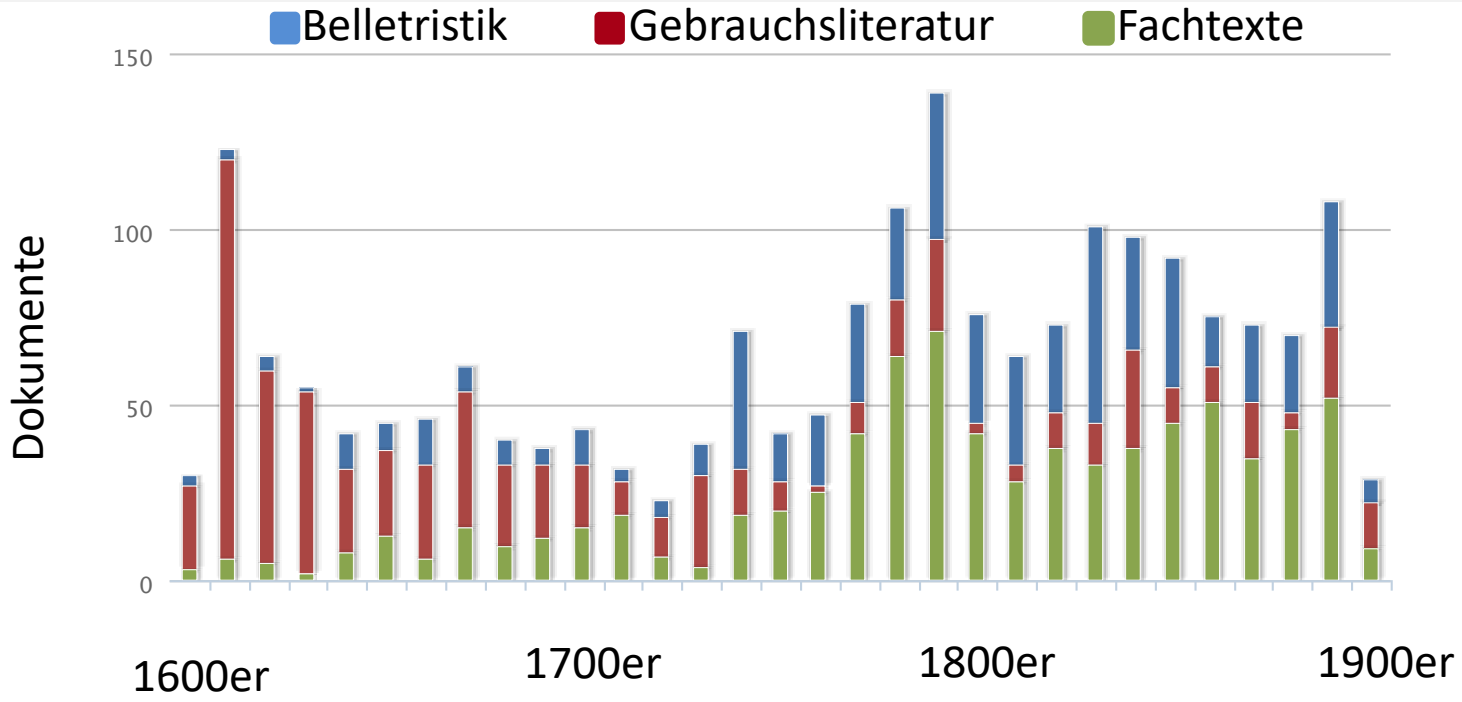


Ressourcen

- Computerlingustik ist stark datengetrieben:
- *Seed*-Lexikon:
  - ANGST (Schmidtke et al. (2014))
  - 1.000 deutsche Wort-Emotionspaare
- Zielkorpus:
  - Deutsches Textarchiv (DTA)
  - viele Meta-Daten verfügbar (z.B. Klassen und Unterklassen)

Lemma	Valence	Arousal	Dominance
Sonnenschein	8.1	5.3	5.4
Terrorismus	1.6	7.4	2.7
Ekstase	6.9	7.8	3.0

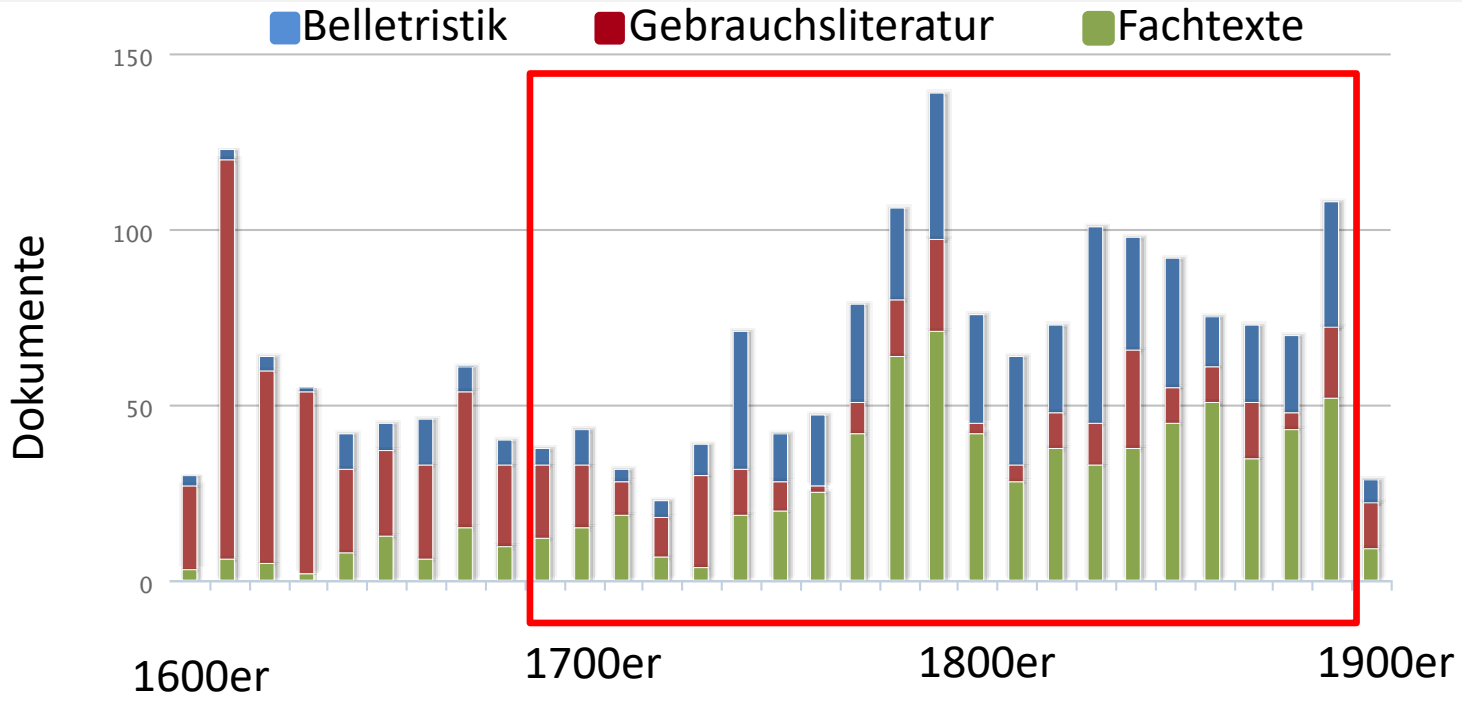
Zielkorpus: DTA



http://www.deutsches-textarchiv.de/doku/textauswahl

- Aggregation von jeweils 30-Jahres-Abschnitten
- Auswahlzeitraum 1690-1899 (~ 1000 Dokumente, 7 Abschnitte)

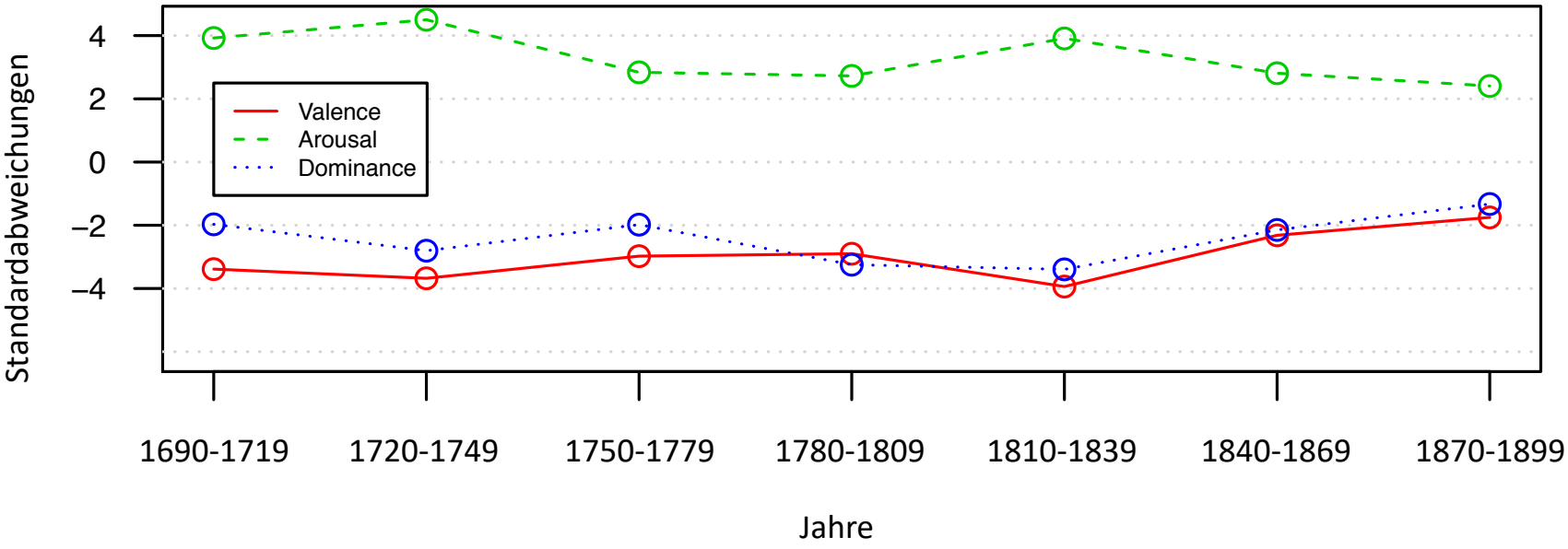
Zielkorpus: DTA



http://www.deutsches-textarchiv.de/doku/textauswahl

- Aggregation von jeweils 30-Jahres-Abschnitten
- Auswahlzeitraum 1690-1899 (~ 1000 Dokumente, 7 Abschnitte)

Emotionsverlauf von *Sünde*



- Starker Anstieg der Valenz seit Beginn der Aufklärung
- Vergleich mit korpuslinguistischen Methoden

Die 10 häufigsten Kollokationen von *Sünde*

Rang	Lemma	
	1690er	1890er
1	todt-	Lamm
2	Erzürnung	hinwegnehmen
3	läßlich	Verzeihung
4	beichten	Ausschweifung
5	Nachlaß	Gotte
6	Grobheit	Schande
7	verschweigen	Reue
8	beweinen	Ärgernis
9	pichen	Laster
10	beichten	aufrichtig

Die 10 häufigsten Kollokationen von *Sünde*

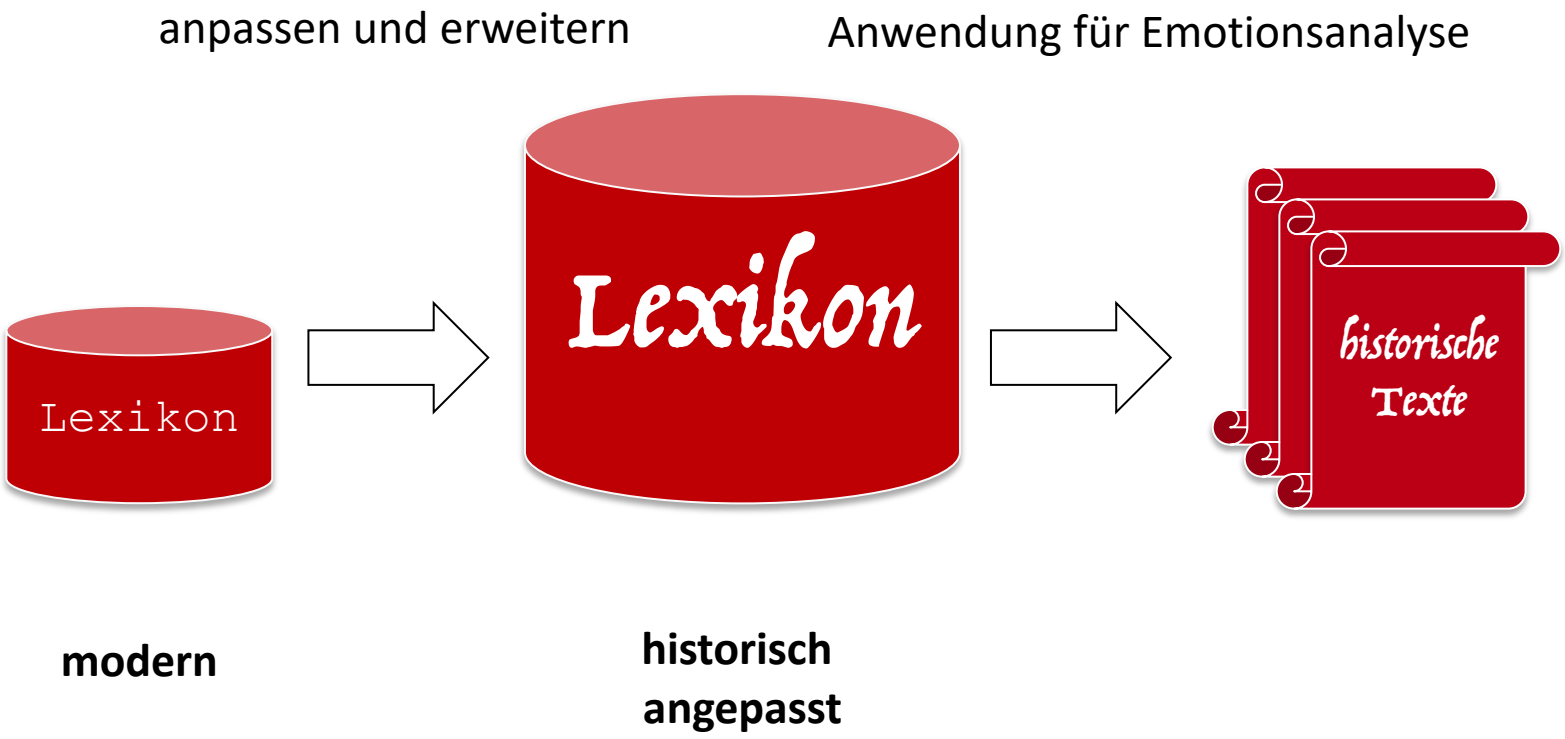
Rang	Lemma	
	1690er	1890er
1	todt-	Lamm
2	Erzürnung	hinwegnehmen
3	läßlich	Verzeihung
4	beichten	Ausschweifung
5	Nachlaß	Gotte
6	Grobheit	Schande
7	verschweigen	Reue
8	beweinen	Ärgernis
9	pichen	Laster
10	beichten	aufrichtig

Die 10 häufigsten Kollokationen von *Sünde*

Rang	Lemma	
	1690er	1890er
1	todt-	Lamm
2	Erzürnung	hinwegnehmen
3	läßlich	Verzeihung
4	beichten	Ausschweifung
5	Nachlaß	Gotte
6	Grobheit	Schande
7	verschweigen	Reue
8	beweinen	Ärgernis
9	pichen	Laster
10	beichten	aufrichtig

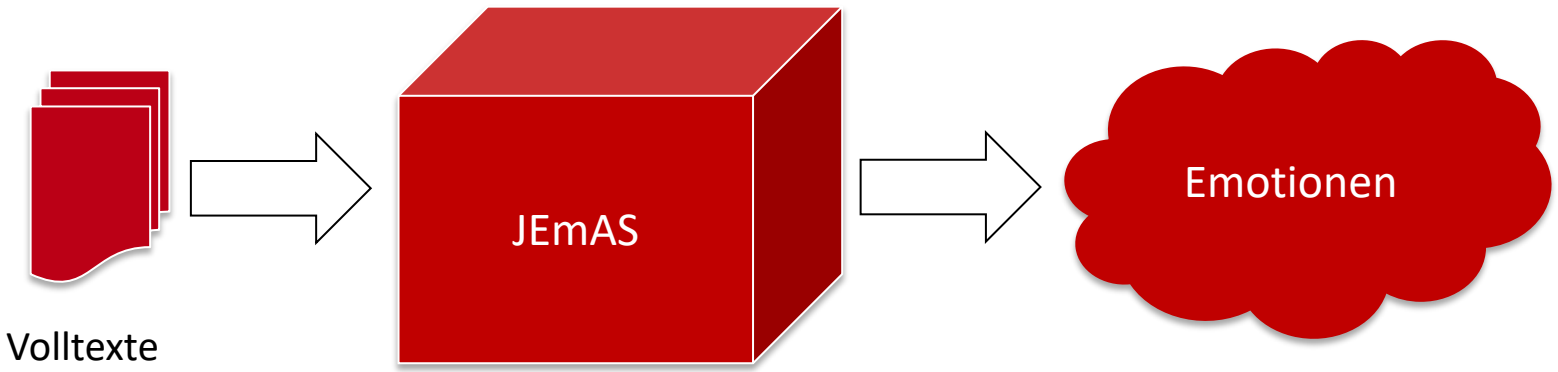
Experiment II
(Quantifizierung historischer Textemotionen)

Methodisches Vorgehen



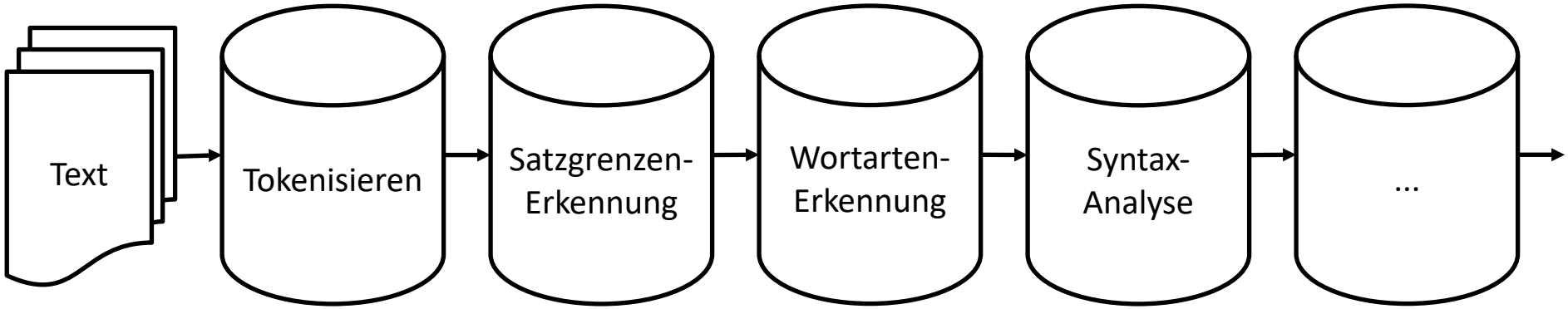
Textbasierte Emotionsanalyse: JEMAS

(Buechel & Hahn, ECAI 2016)



Verfügbar:
`https://github.com/JULIELab/JEmAS`

Pipeline-Modell der Computerlinguistik



- Reihung voneinander unabhängiger Komponenten
- Schrittweise Analyse unter Rückgriff früherer Ergebnisse

Lexikonbasierter Ansatz

Kennst du das Land, wo die Zitronen blühen?

- Tokenisieren

  Kennst | du | das | Land | , | wo | die | Zitronen | blühen | ?

- Lemmatisieren

  Kennen | du | der | Land | , | wo | der | Zitrone | blühen | ?

- Stoppwörter und Interpunktion entfernen

  Kennen | _ | _ | Land | _ | _ | _ | Zitrone | blühen | _

- Lexikonabgleich

  (6,5,5) | (6,5,5) | (6,6,4) | (8,7,7)

- VAD-Berechnung

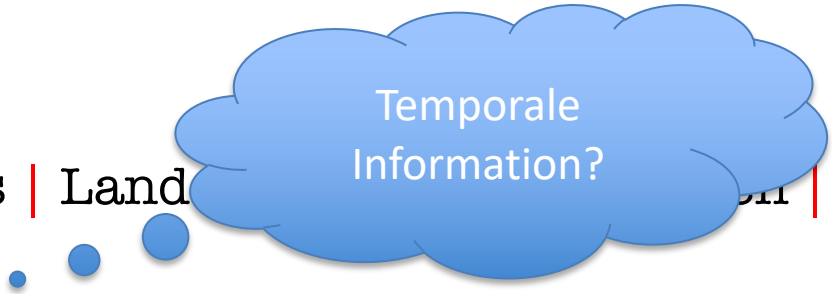
  [(6,5,5) + (6,5,5) + (6,6,4) + (8,7,7)] / 4 = (6.5, 5.75, 5.25)

Lexikonbasierter Ansatz

Kennst du das Land, wo die Zitronen blühen?

- Tokenisieren

  Kennst | du | das | Land | … | blühen | ?



- Lemmatisieren

  Kennen | du | der | Land | , | wo | der | Zitrone | blühen | ?

- Stoppwörter und Interpunktion entfernen

  Kennen | _ | _ | Land | _ | _ | _ | Zitrone | blühen | _

- Lexikonabgleich

  (6,5,5) | (6,5,5) | (6,6,4) | (8,7,7)

- VAD-Berechnung

  [(6,5,5) + (6,5,5) + (6,6,4) + (8,7,7)] / 4 = (6.5, 5.75, 5.25)

Lexikonbasierter Ansatz

Kennst du das Land, wo die Zitronen blühen?

- Tokenisieren

  Kennst | du | das | Land | … | blühen | ?

- Lemmatisieren

  Kennen | du | der | Land | , | wo | der | Zi…

- Stoppwörter und Interpunktion entfernen

  Kennen | _ | _ | Land | _ | _ | _ | Zitrone | blühen | _

- Lexikonabgleich

  (6,5,5) | (6,5,5) | (6,6,4) | (8,7,7)

- VAD-Berechnung

  [(6,5,5) + (6,5,5) + (6,6,4) + (8,7,7)] / 4 = (6.5, 5.75, 5.25)

Temporale Information?

Negation??

Lexikonbasierter Ansatz

Kennst du das Land, wo die Zitronen blühen?

- Tokenisieren

  Kennst | du | das | Land | … | blühen | ?

- Lemmatisieren

  Kennen | du | der | Land | , | wo | der | Zi…

- Stoppwörter und Interpunktion entfernen

  Kennen | _ | _ | Land | _ | _ | _ | Zitrone | blühen | _

- Lexikonabgleich

  (6,5,5) | (6,5,5) | (6,6,4) | (8,7,7)

- VAD-Berechnung

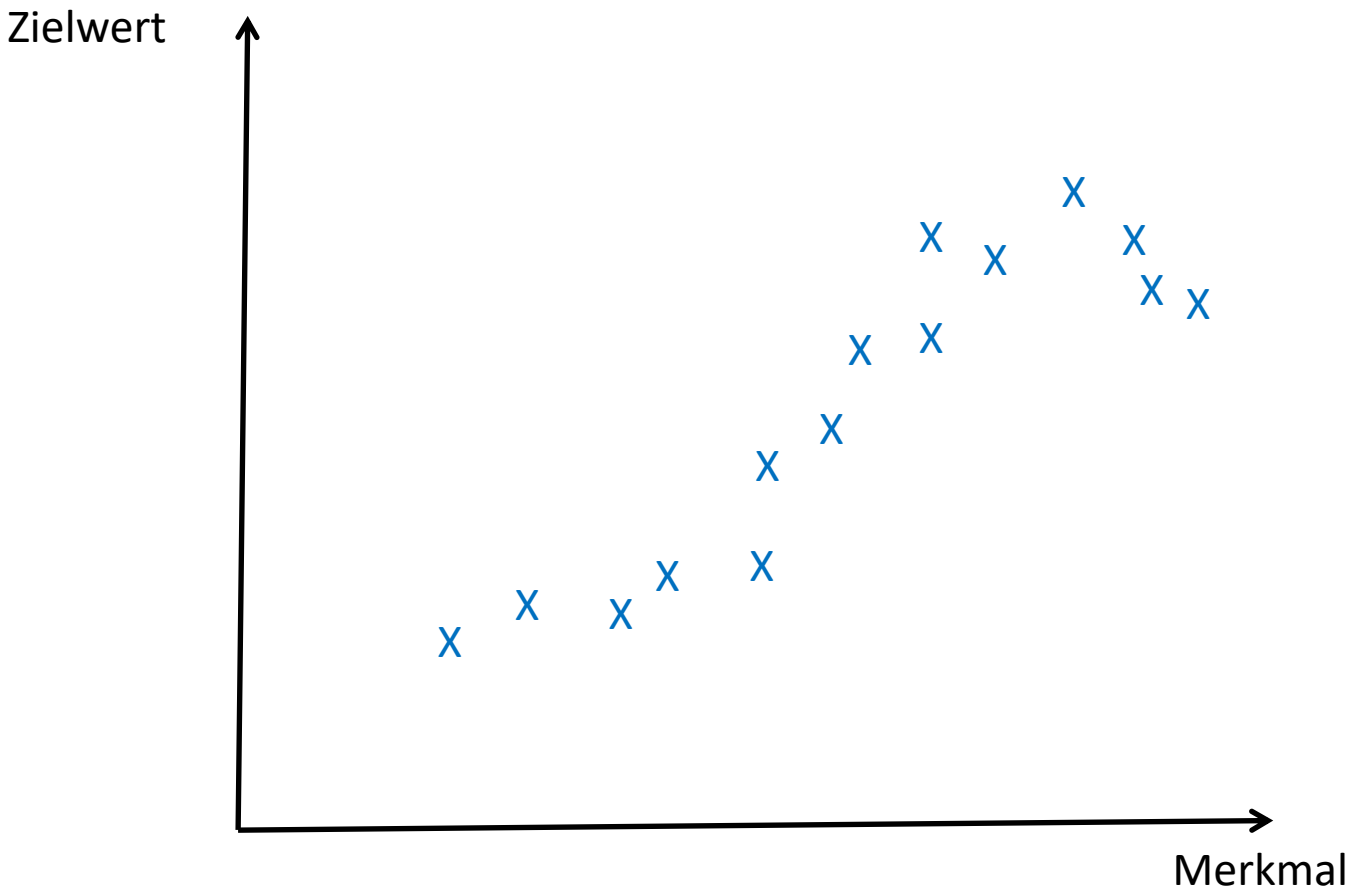
  [(6,5,5) + (6,5,5) + (6,6,4) + (8,7,7)] / 4 = (6.5, 5.75, 5.25)

Temporale Information?

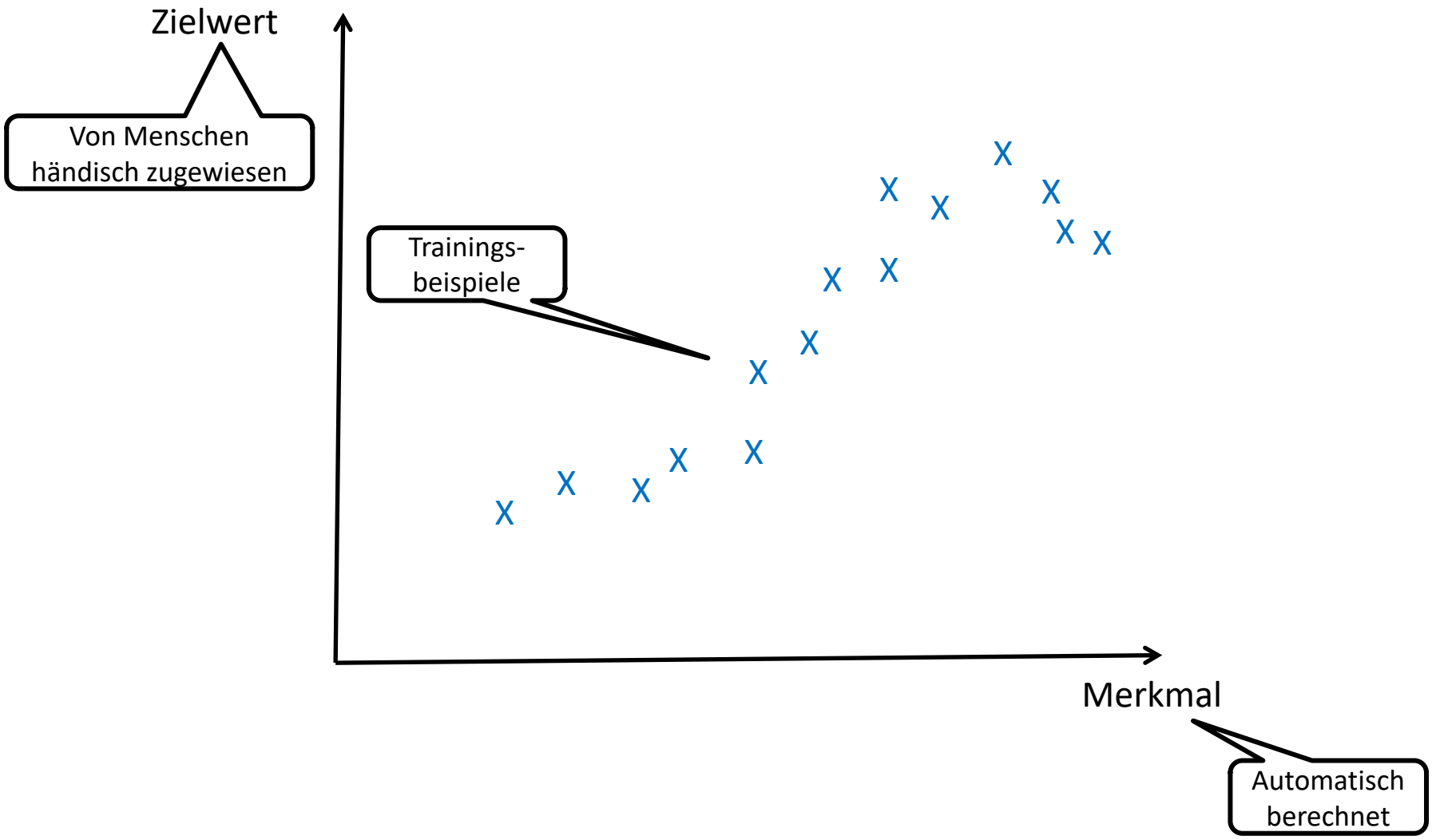
Negation??

Sequenzielle Ordnung???

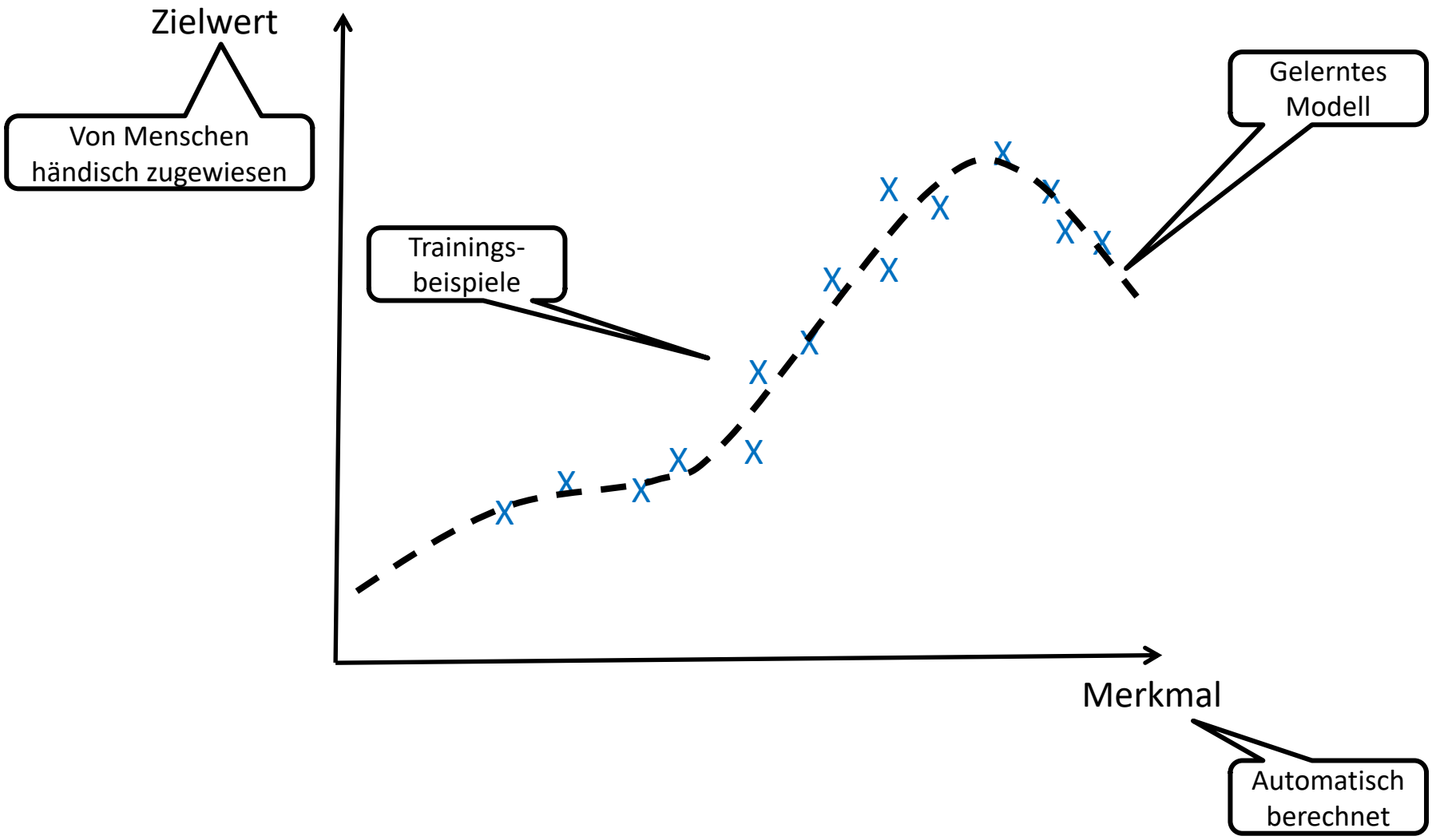
Maschinenlernen



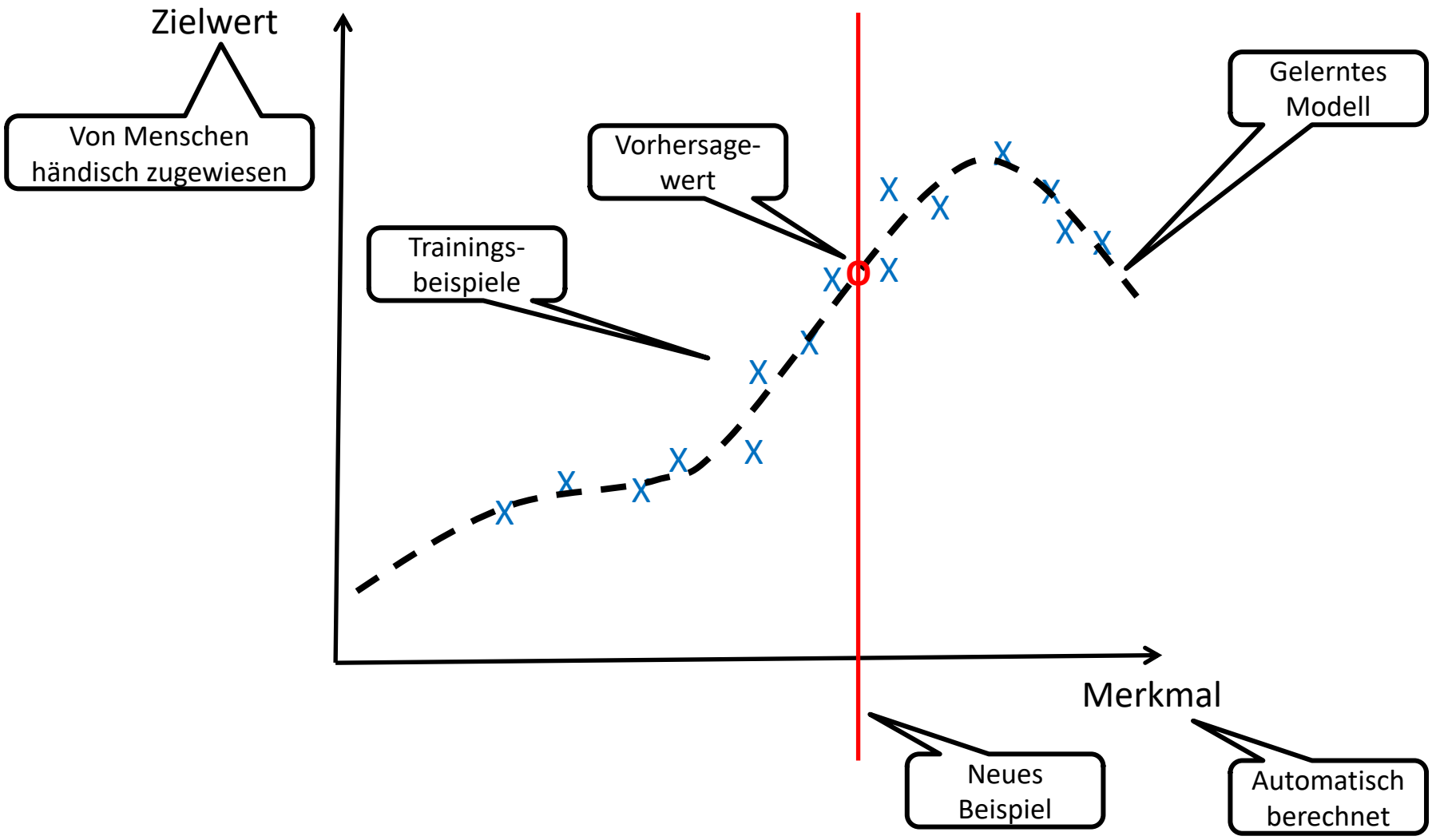
Maschinenlernen



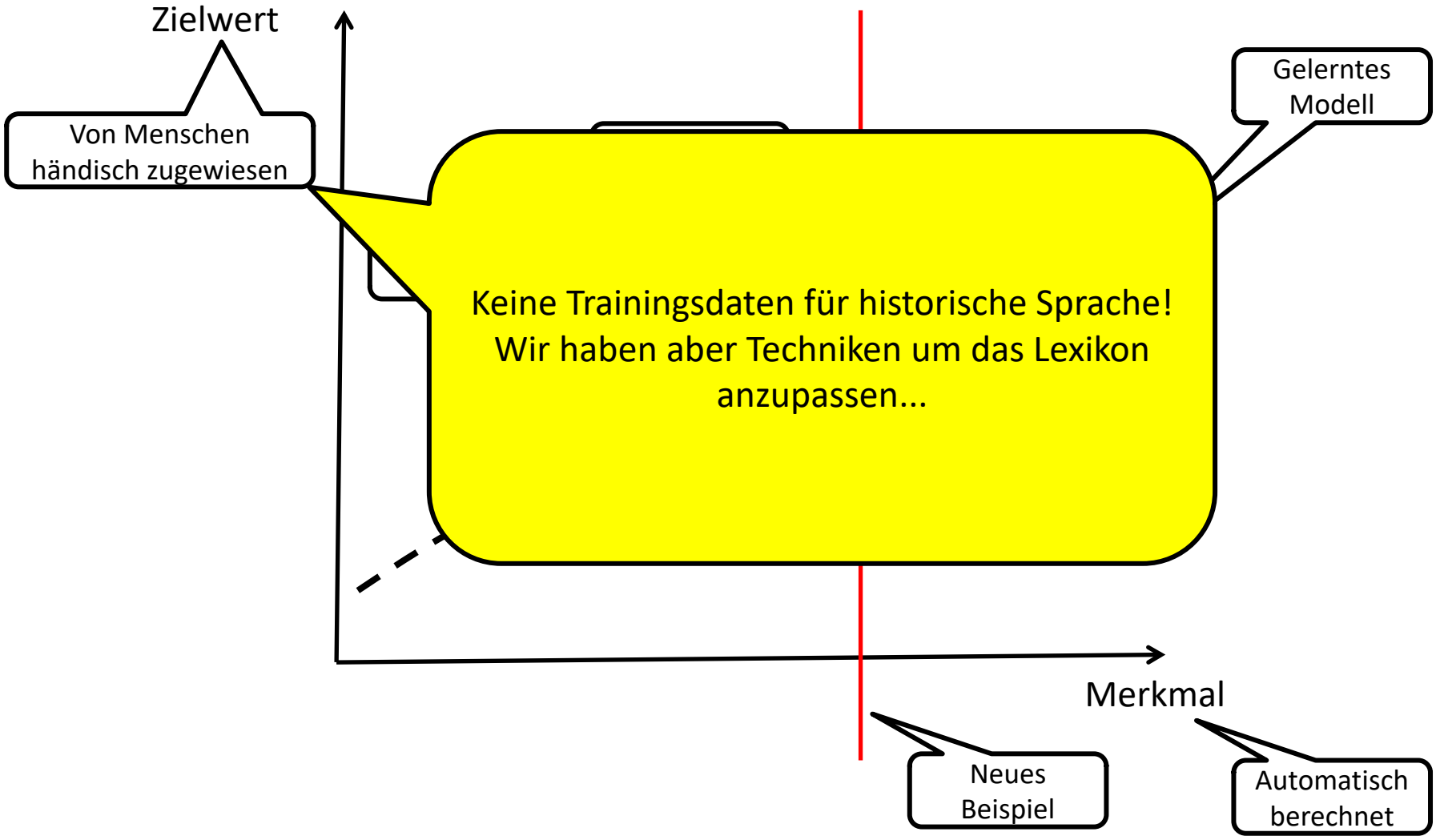
Maschinenlernen



Maschinenlernen

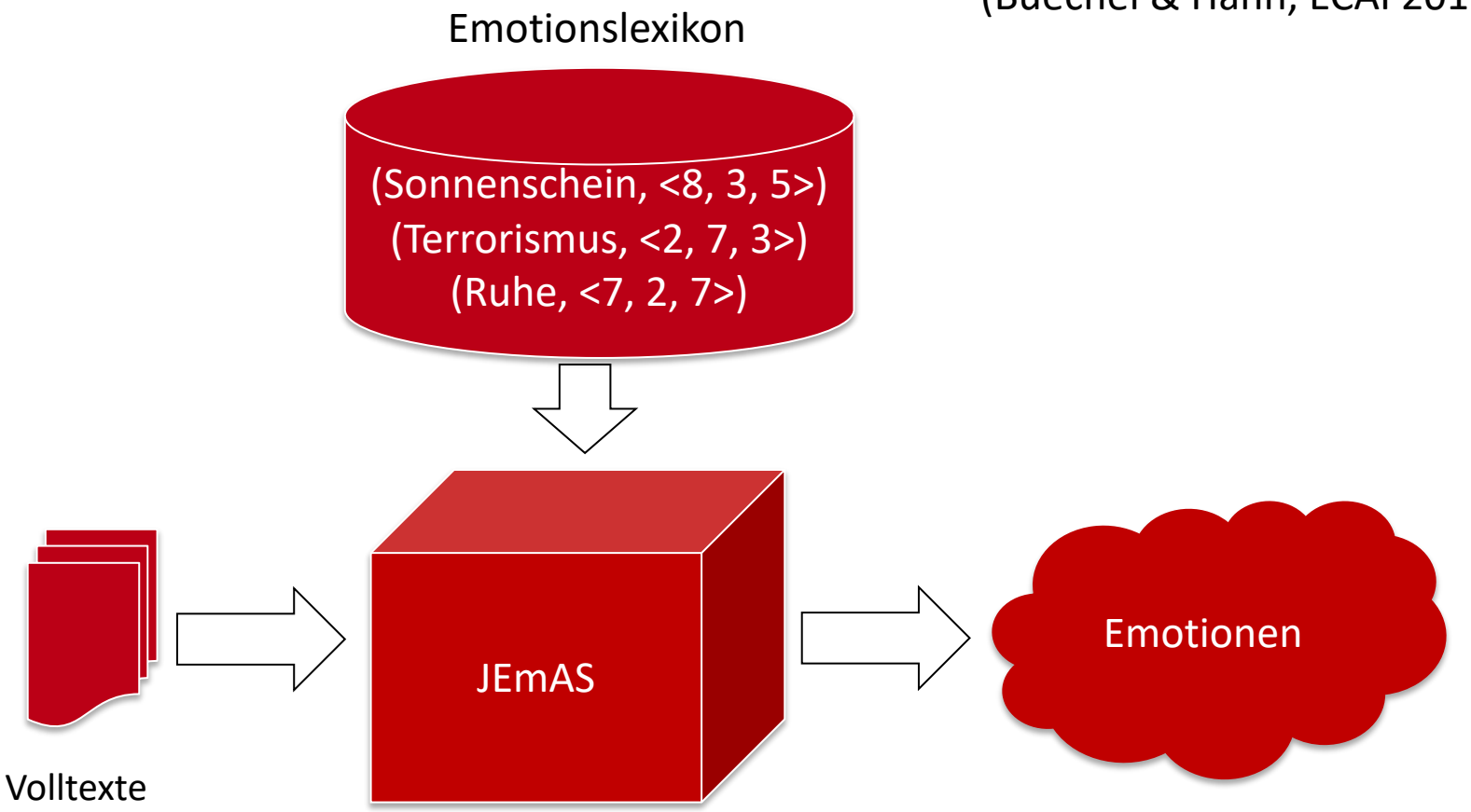


Maschinenlernen



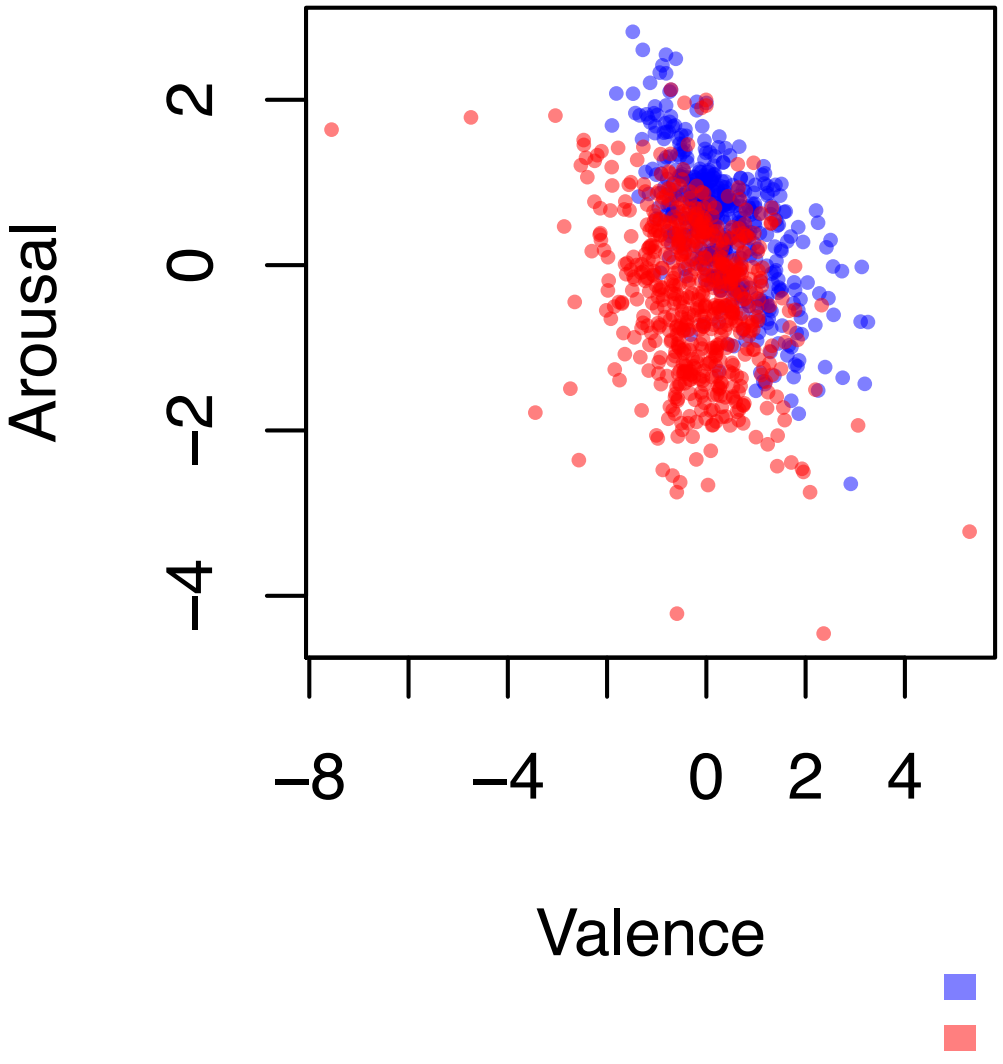
Textbasierte Emotionsanalyse: JEmAS

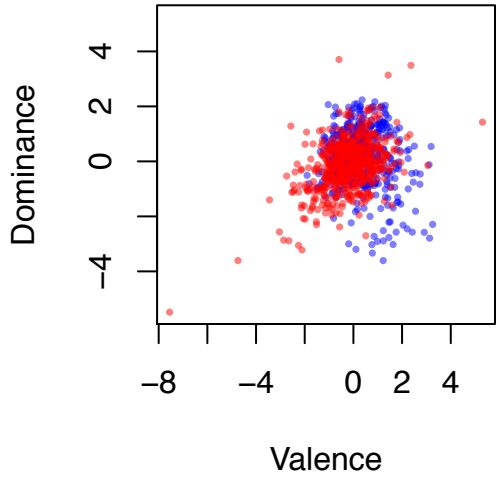
(Buechel & Hahn, ECAI 2016)

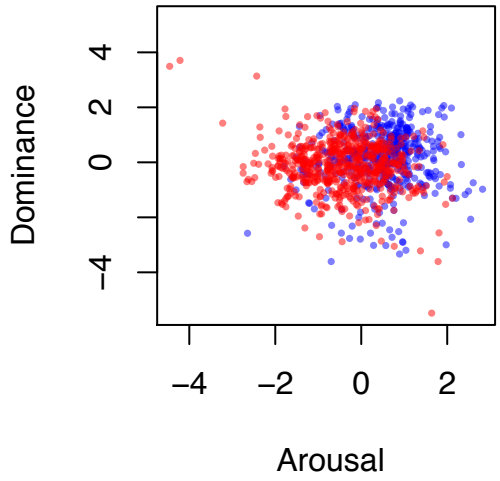


Verfügbar:
`https://github.com/JULIELab/JEmAS`

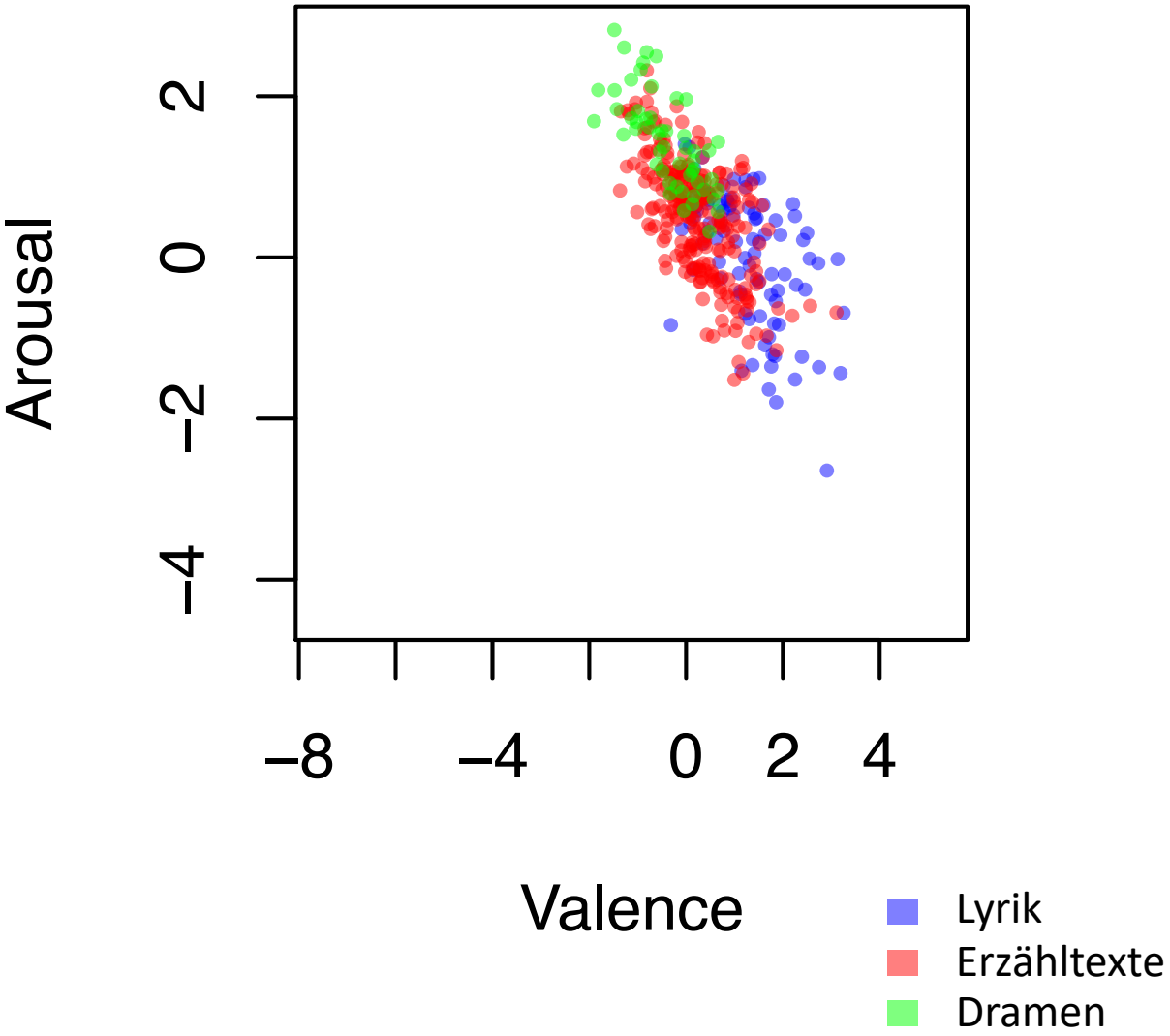
Unterscheidung der DTA-Textsorten

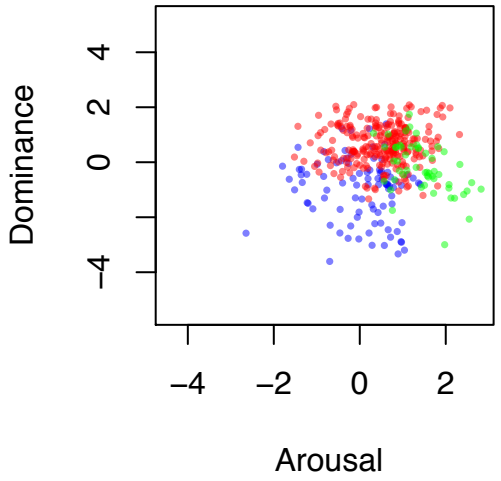


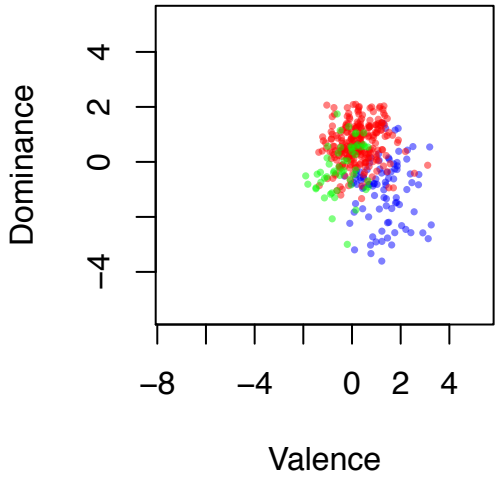




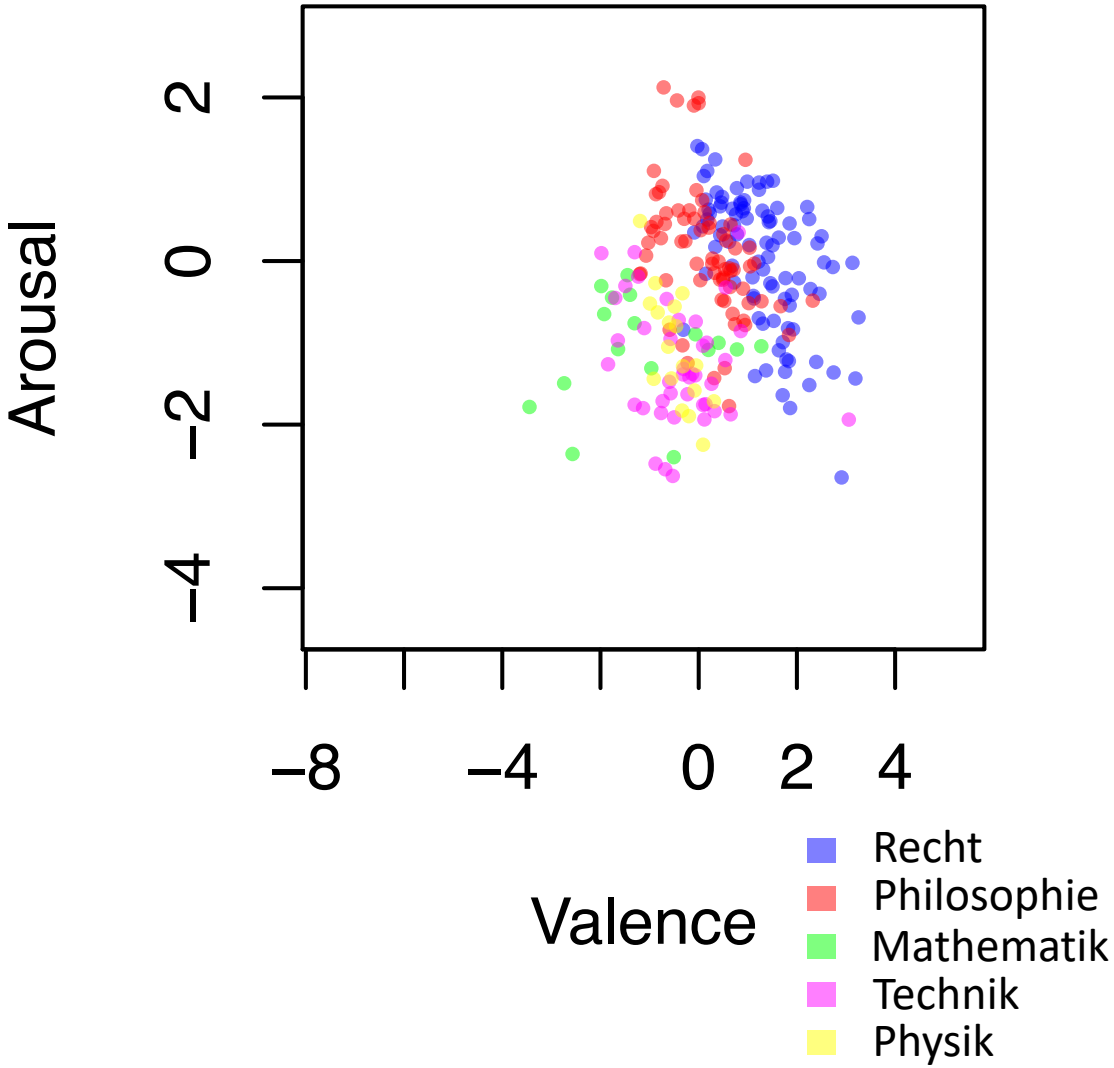
Unterscheidung literarischer Gattungen

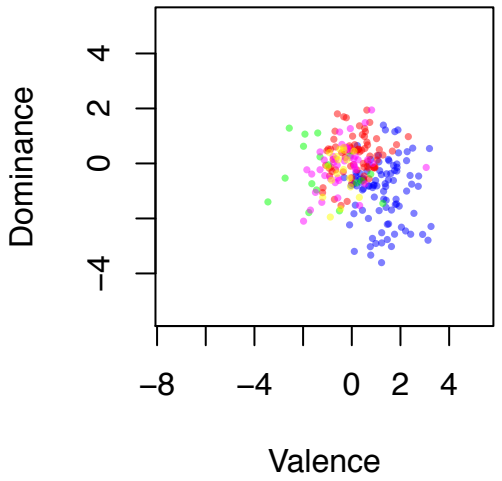


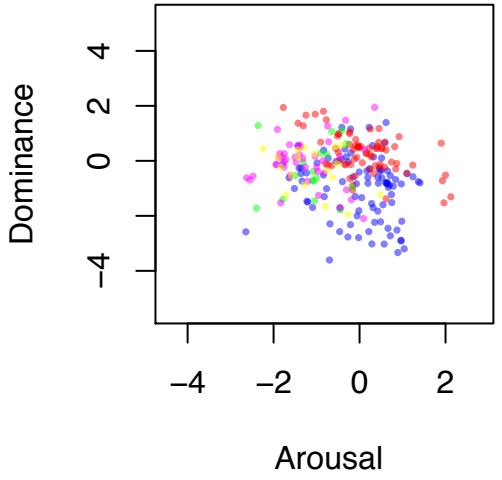




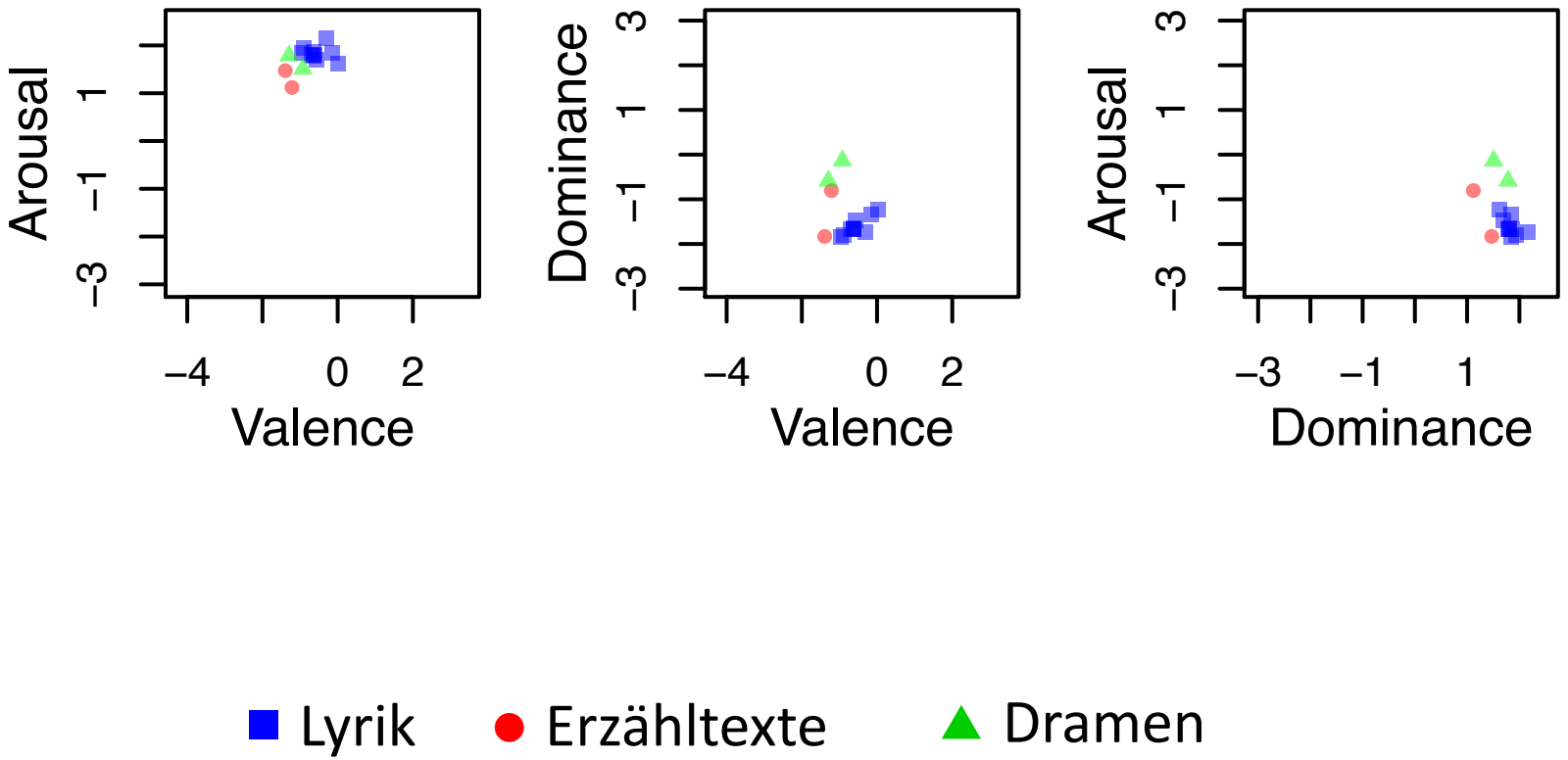
Unterscheidung akademischer Disziplinen



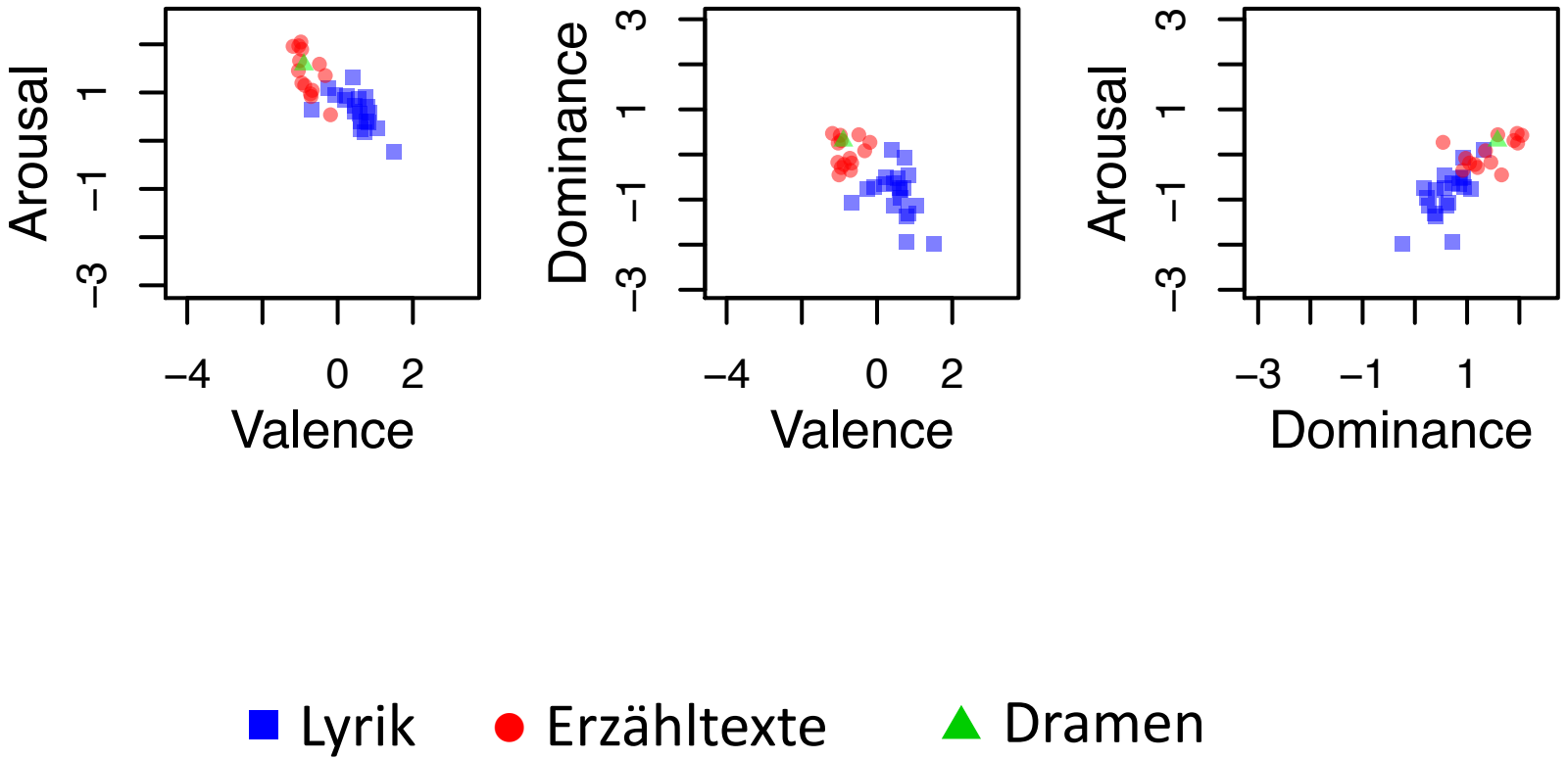




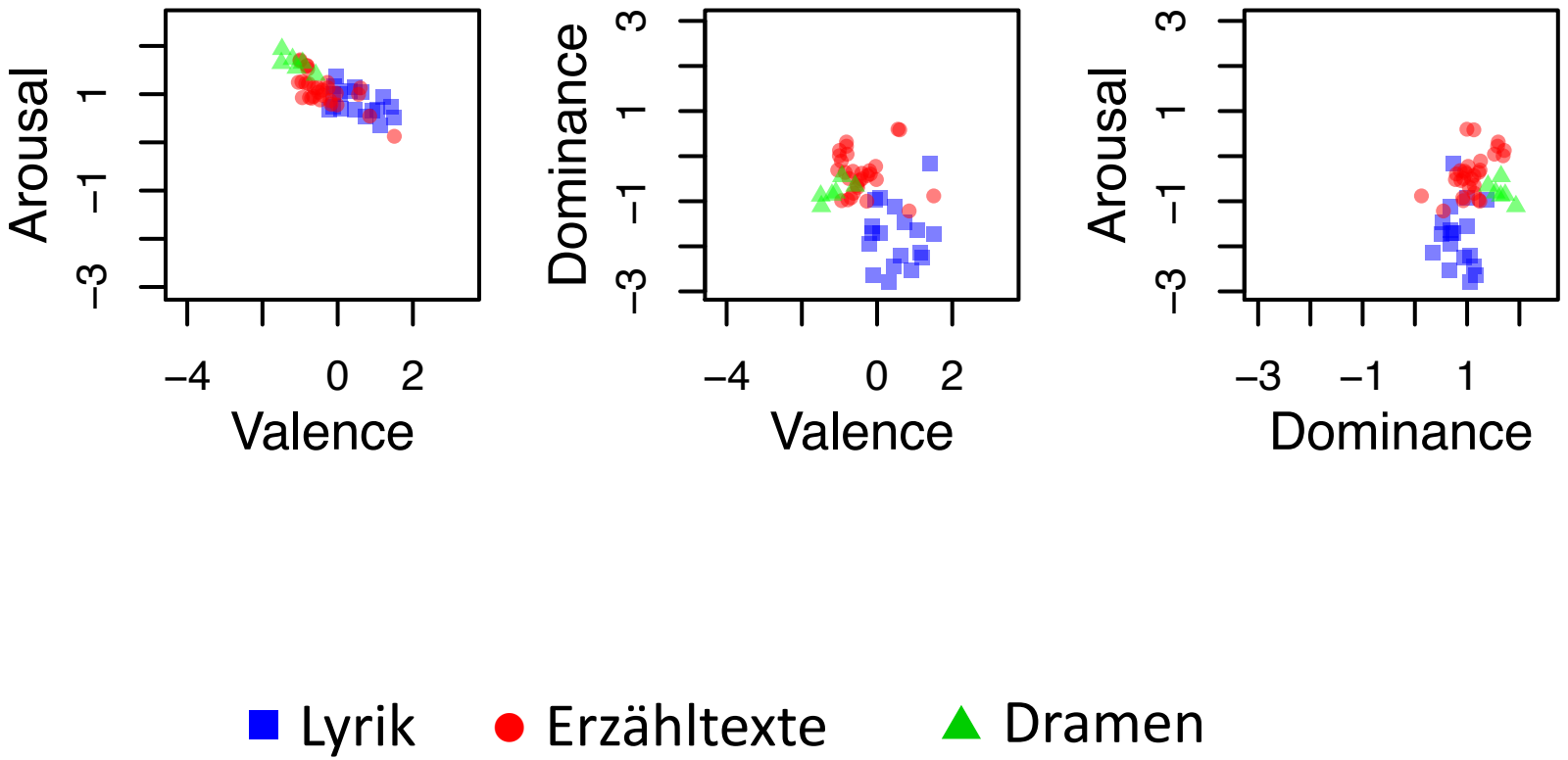
Entwicklung literarischer Gattungen (1690-1719)



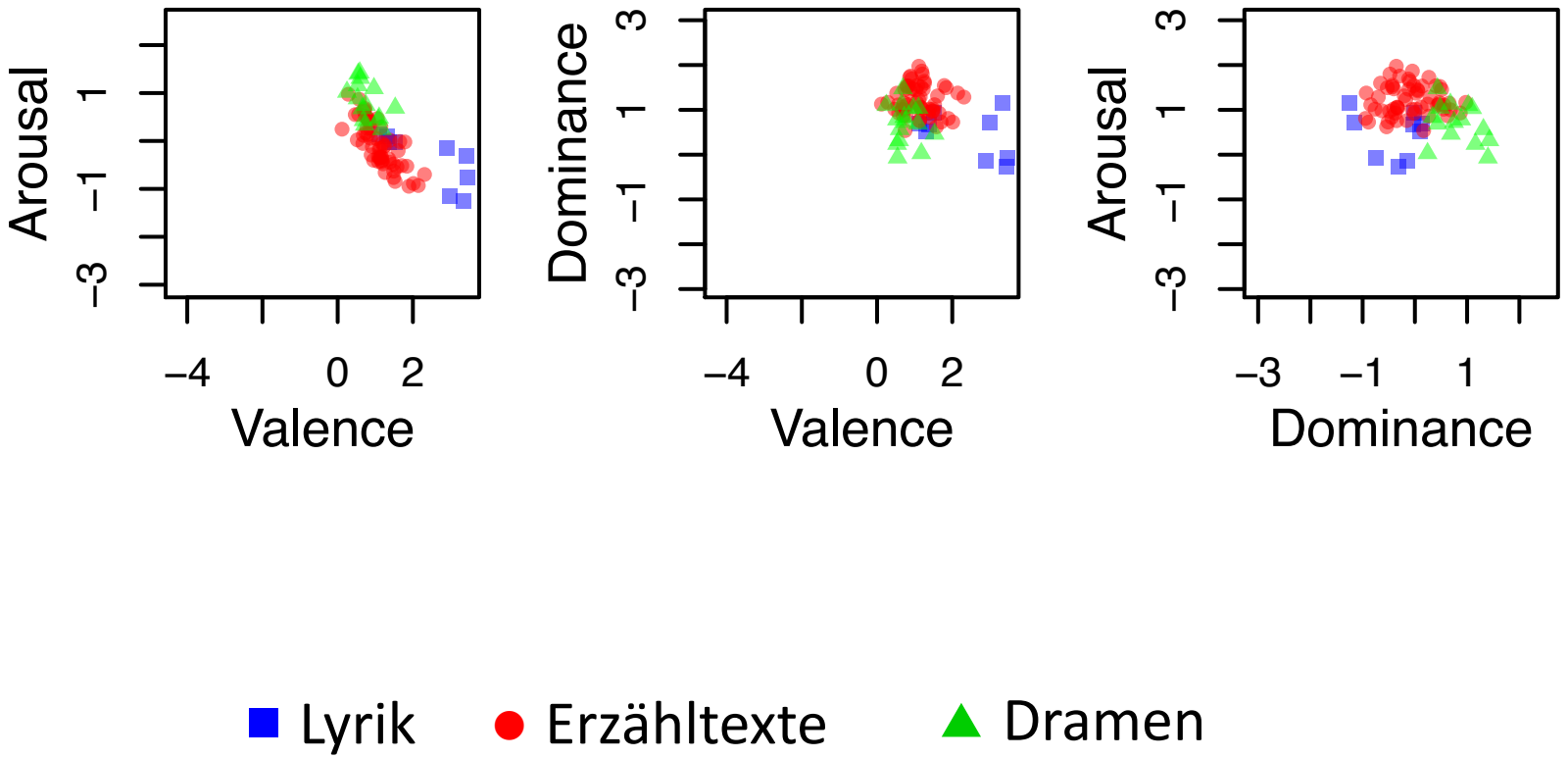
Entwicklung literarischer Gattungen (1720-1749)



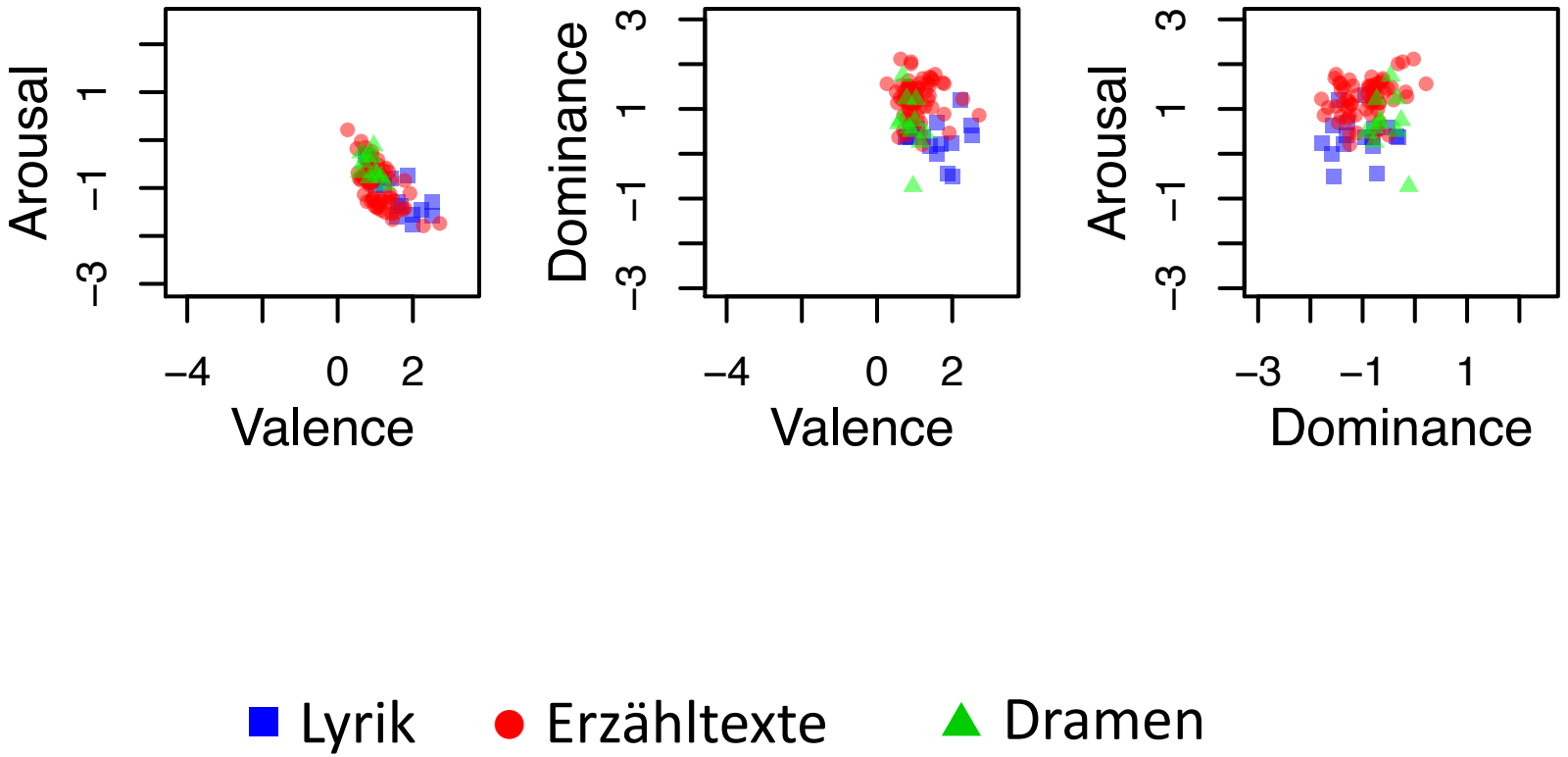
Entwicklung literarischer Gattungen (1750-1779)



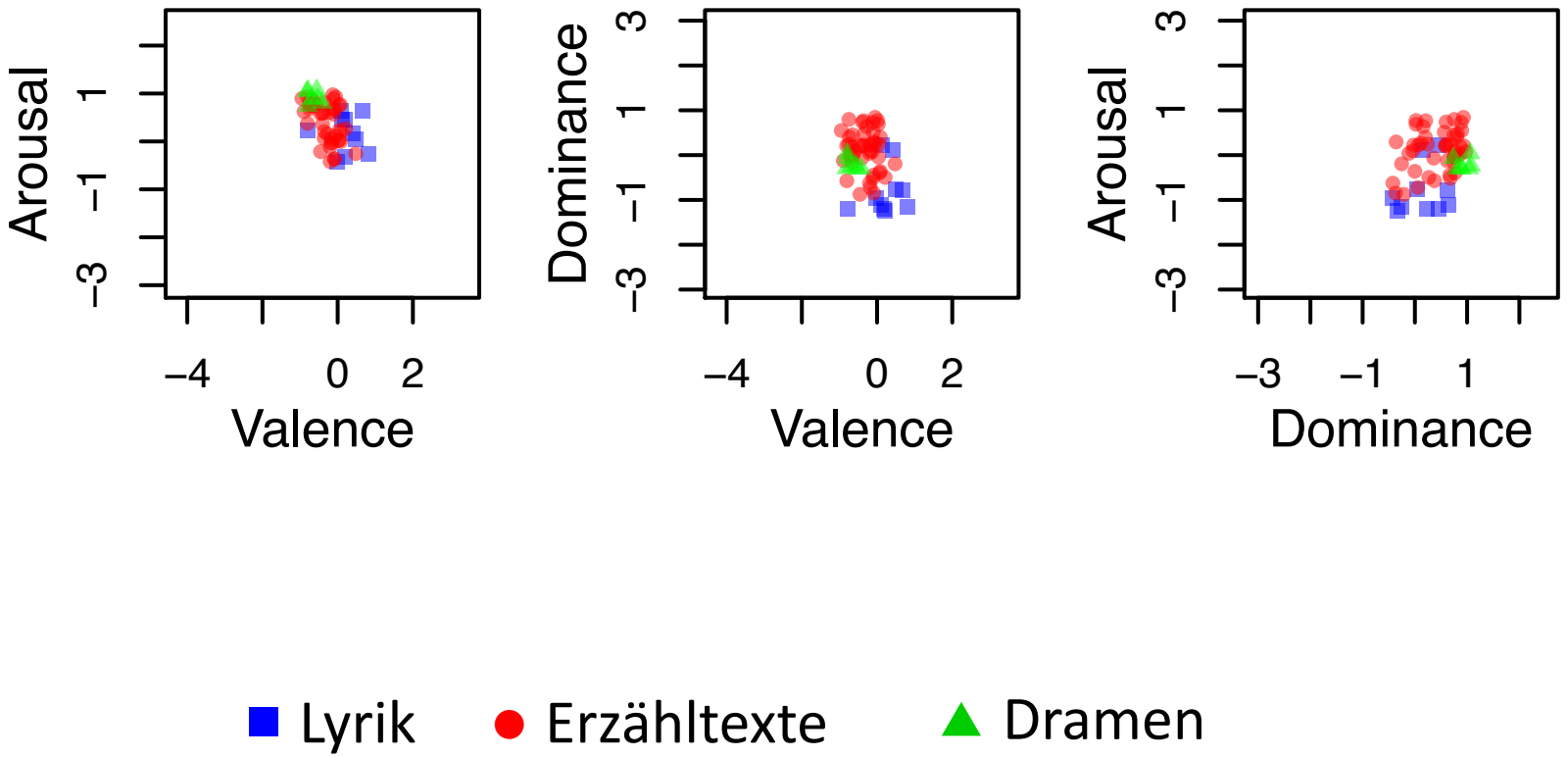
Entwicklung literarischer Gattungen (1780-1809)



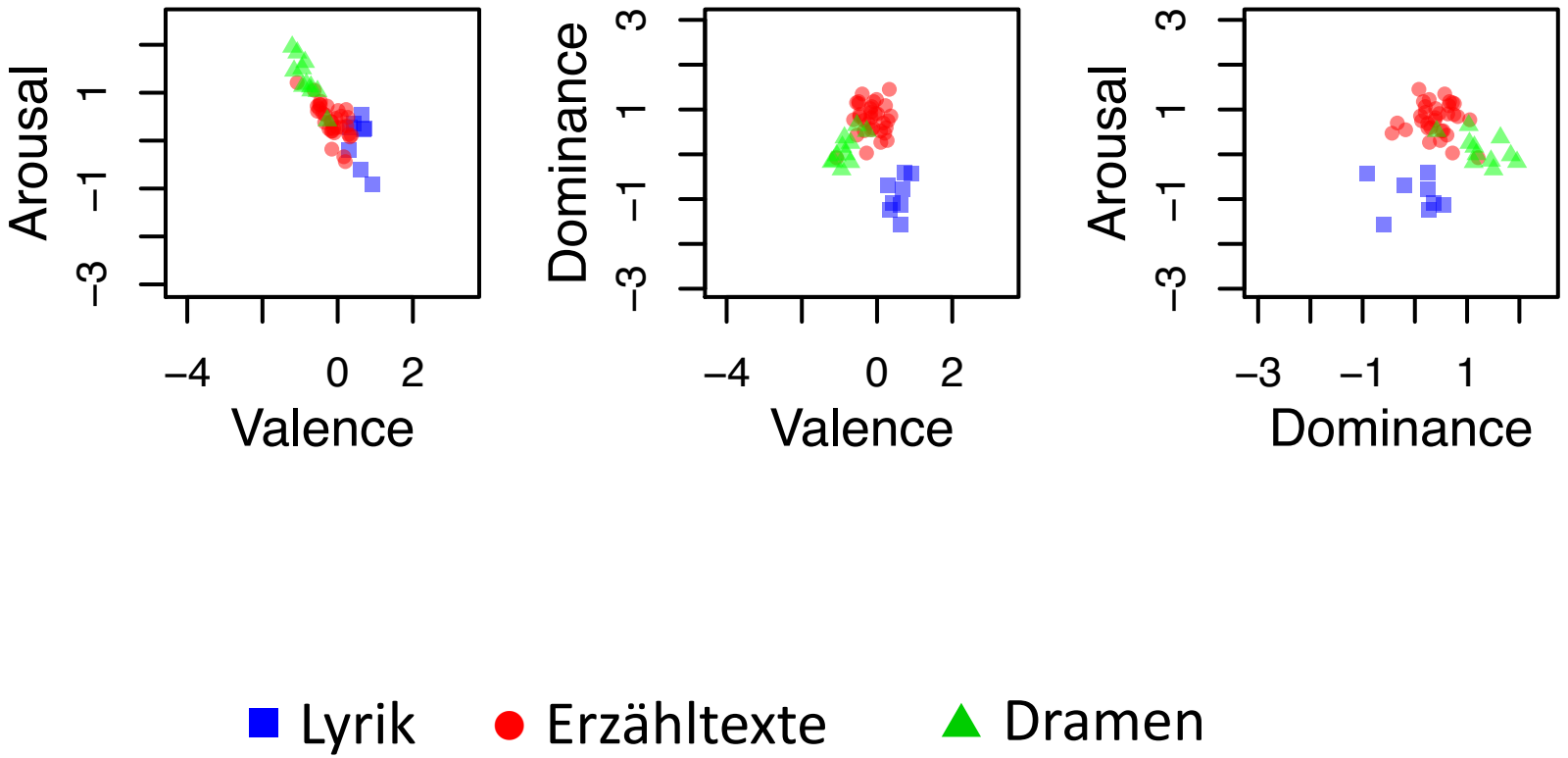
Entwicklung literarischer Gattungen (1810-1839)



Entwicklung literarischer Gattungen (1840-1869)



Entwicklung literarischer Gattungen (1870-1899)



Zusammenfassung

Zusammenfassung

- Methode für die Quantifizierung von Emotionen in historischer Sprache
- Kombination aus Sprachstufenanpassung für Emotionslexika und lexikonbasierter Textanalyse
- Fallstudie zur Anwendung auf das DTA-Korpus
- Qualitative Validierung auf Wort- und Textebene

➢ Gattungen zeigen distinktive, zeitlich veränderliche Emotionsprofile

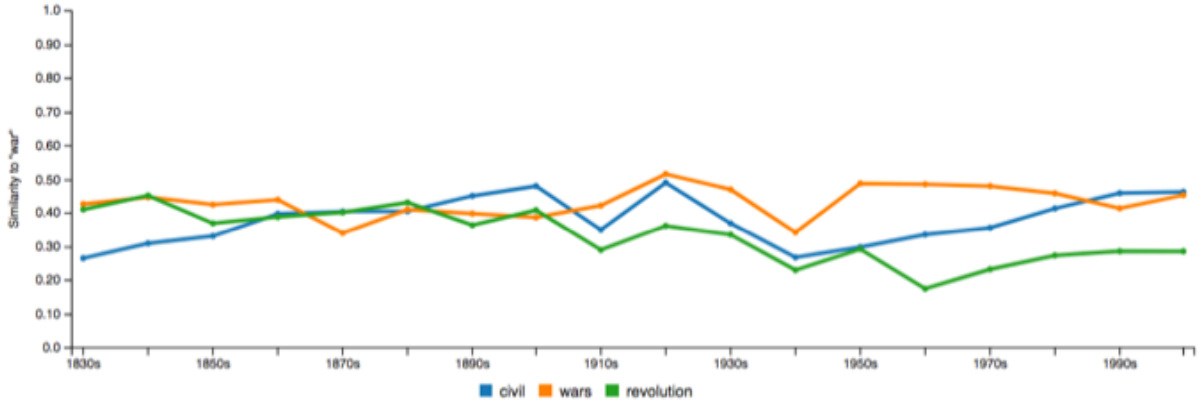
➢ Ressourcen verfügbar: `https://github.com/JULIELab/HistEmo`

➢ Webportal zum Abfragen der Emotionswerte: `http://jeseme.org`

JeSemE 2.0 - The Jena Semantic Explorer

Results for "war" in Corpus of Historical American English
Note: lowercased
Search in Corpus of Historical American English

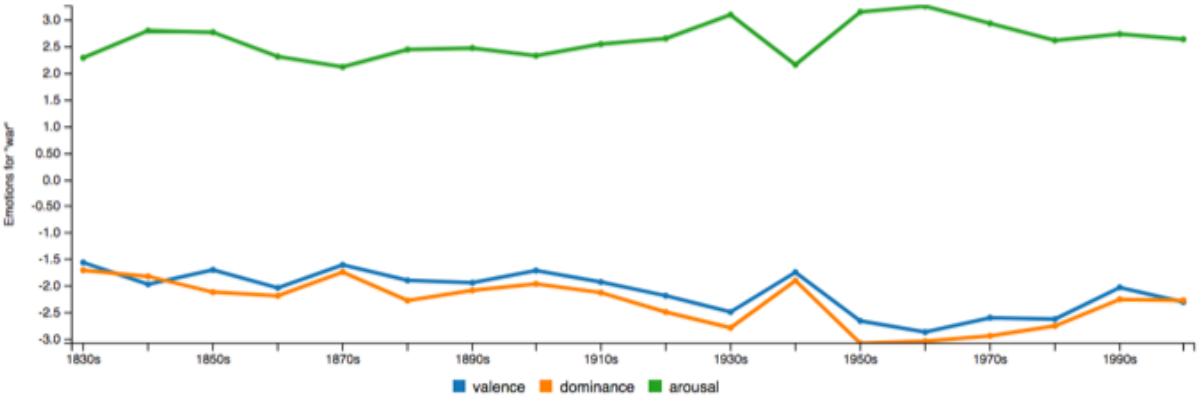
Similar Words



This chart shows the words detected as most similar to "war" and how their similarity changed over time (higher is more similar). These values are based on SVD_{PPMI} — see Help for details.

Add word to graph: add

Word Emotion



This chart shows how the emotions connected to "war" changed over time. We describe emotions via valence (positive – negative), arousal (calm – excited) and dominance (controlled – in control) — see Help for details. Scores are given in standard deviations around an (for each time span) emotionally neutral 0.

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Jena University Language & Information Engineering Lab

Quantifizierung von Emotionen in Historischer Sprache

Anwendungen und Methodische Grundlagen

Sven Büchel

Jena University Language & Information Engineering
(JULIE) Lab
http://www.julielab.de
Friedrich-Schiller-Universität Jena