INSTITUT FÜR DEUTSCHE SPRACHE

Peter Fankhauser, Abteilung Zentrale Forschung

WIE WÖRTER WANDERN - VISUALISIERUNG VON DIACHRONEM WORTGEBRAUCH

Gemeinsam mit Marc Kupietz and Elke Teich

Mitglied der Leibniz-Gemeinschaft

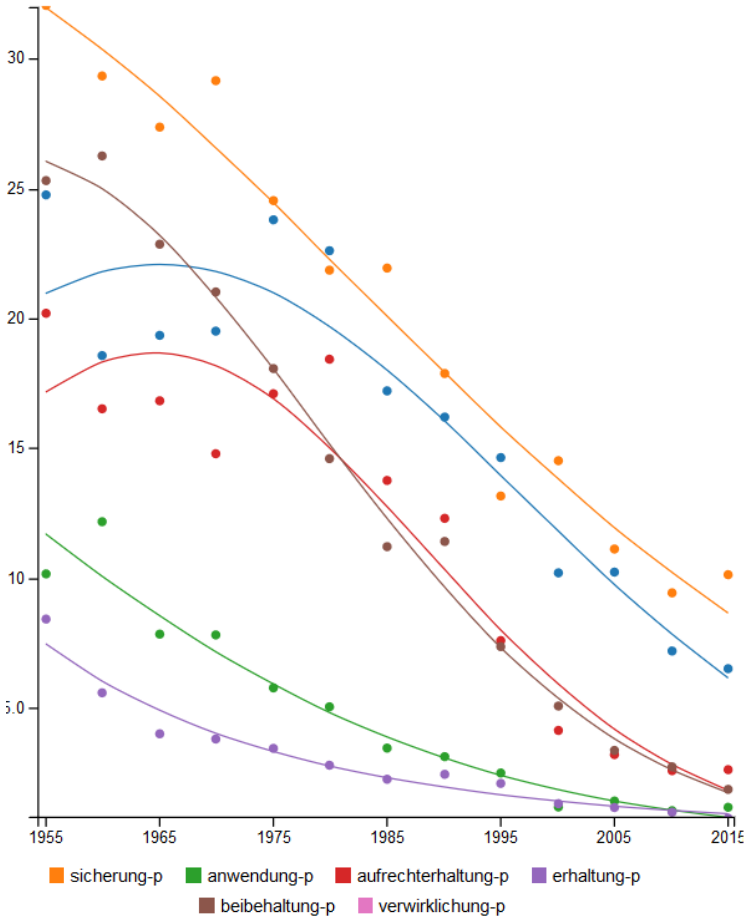
ÜBERSICHT

- Ziel: Exploration von paradigmatischem Wortgebrauchswandel
  - Wörter mit ähnlichem Gebrauch
  - steigen oder fallen gemeinsam
- Ansatz: Visuelle Korrelation von
  - Frequenzänderung
  - (Distributionieller) Semantik
- Beispiele
- Abschließende Bemerkungen

Die Zeit ändert alles;
es gibt keinen Grund, warum
die Sprache diesem allgemeinen
Gesetz enthoben sein sollte.

Ferdinand de Saussure,
Cours de linguistique générale
1916/1959

BEISPIEL: RÜCKGANG VON NOMINALISIERUNG MIT „-UNG“



Sicherung
Anwendung
Aufrechterhaltung
Erhaltung
Beibehaltung
Verwirklichung
(…)

VISUALISIERUNG VON FREQUENZENTWICKLUNG MITTELS FARBE

- Anpassung logistischer Wachstumskurven an Frequenzen *p(t)*

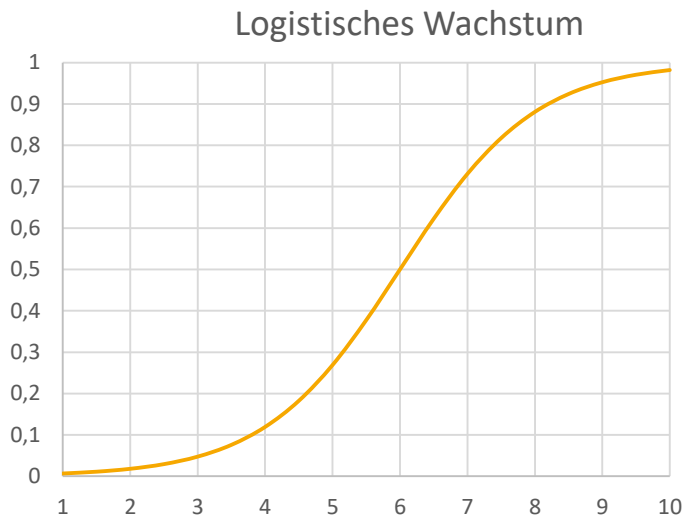
$$p(t) = \frac{1}{1 + e^{-k-s*t}}$$

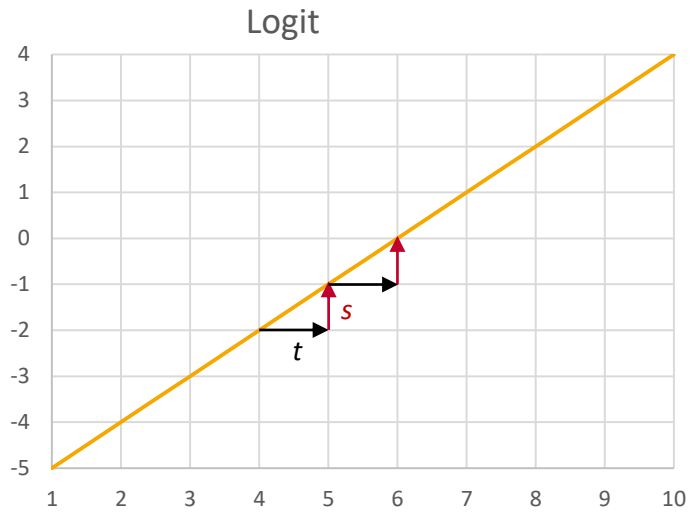
- *k* … Intercept
  s … Steigung
  t … Zeit
- Äquivalent dazu: Logit

$$\ln\left(\frac{p(t)}{1-p(t)}\right) = k + s * t$$

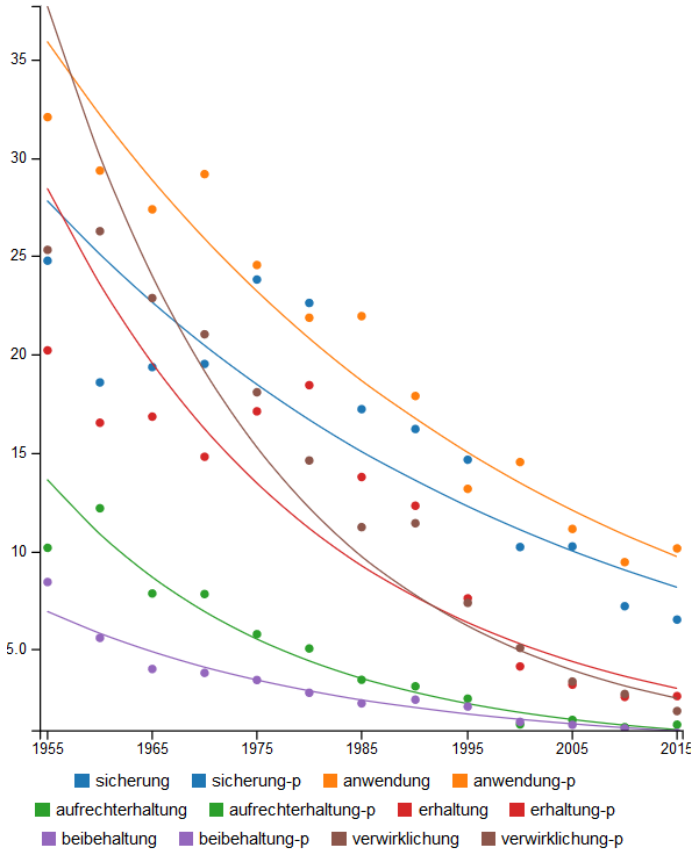
- Abbildung Steigung *s* auf Farbskala

Ähnliche Steigung: Gleiche Farbe

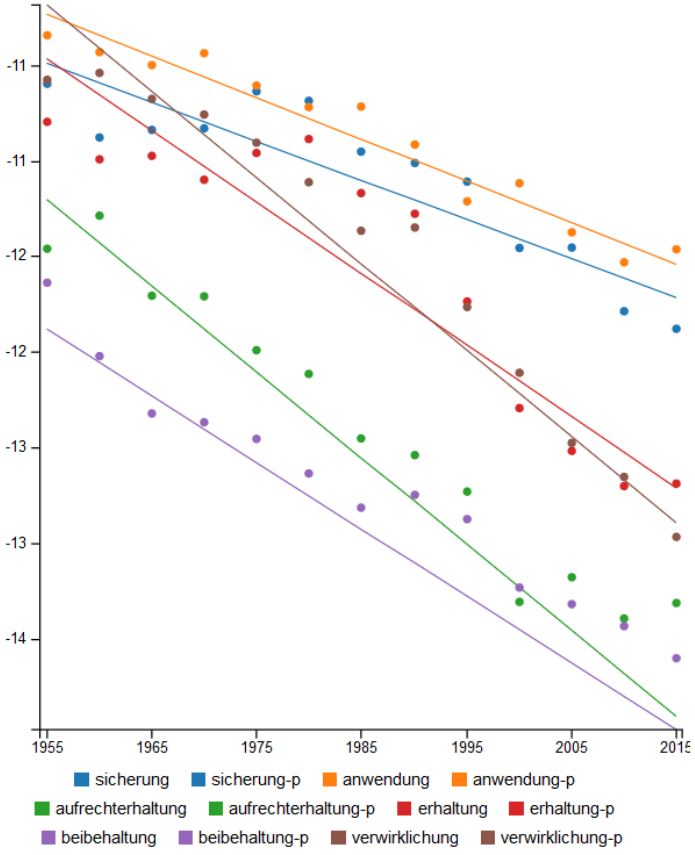




BEISPIEL: FREQUENZEN UND ANGEPASSTE KURVEN

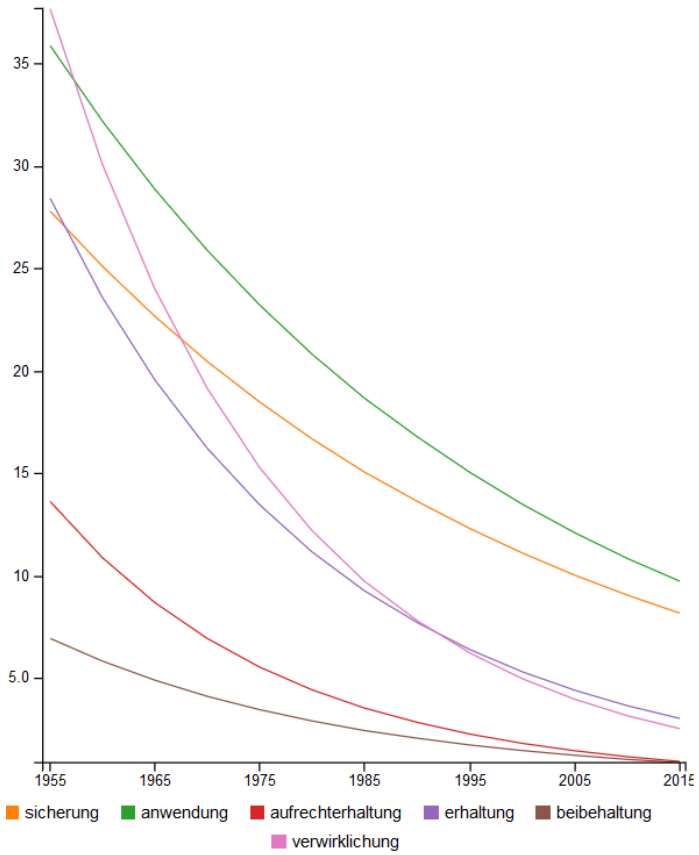


Freq per Mio

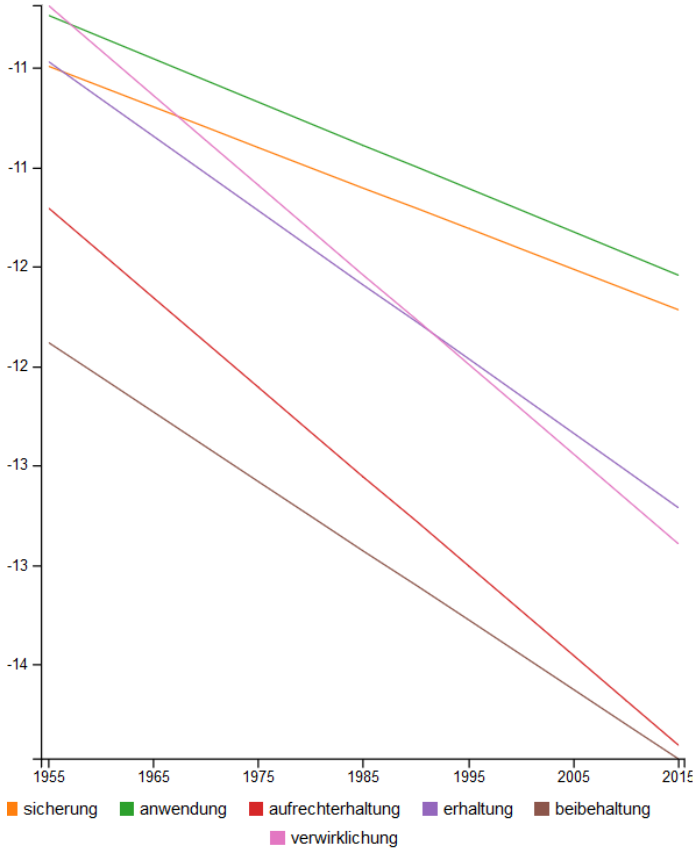


Logit(FpM)

BEISPIEL: ANGEPASSTE KURVEN

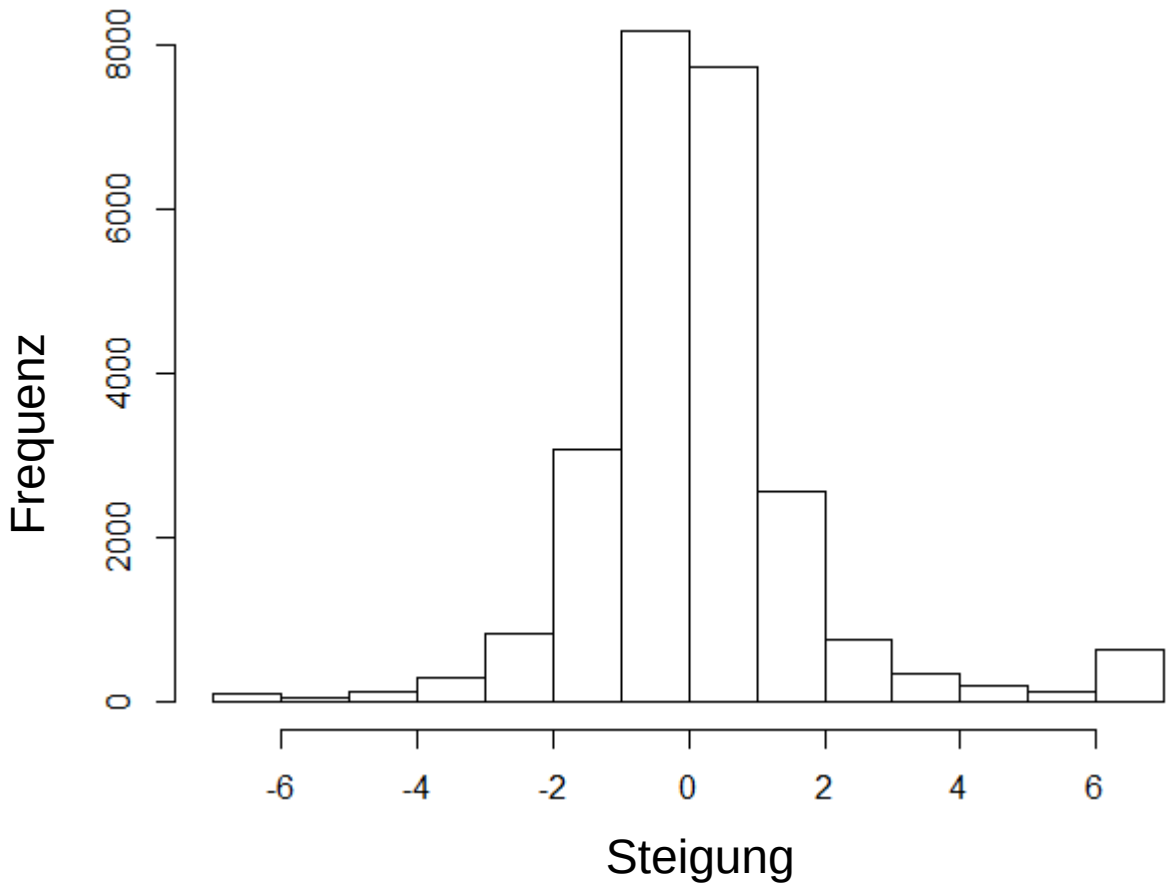


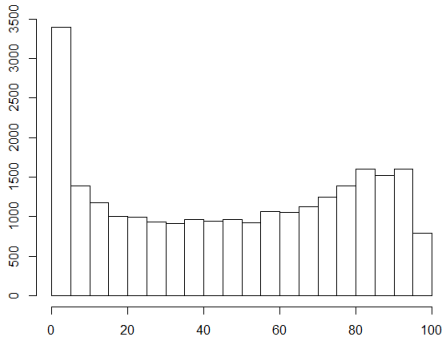
Freq per Mio



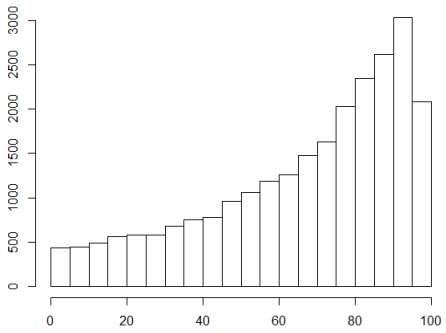
Logit(FpM)

SPIEGEL/ZEIT: VERTEILUNG VON STEIGUNG, PSEUDO-R2





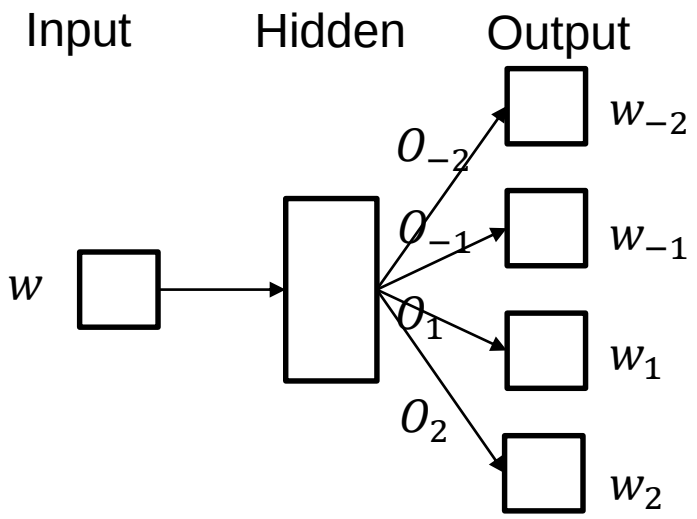
R2 (1): 47.6%



R2 (2): 66.6%

REPRESENTATION VON WORTGEBRAUCH IN WENIGEN DIMENSIONEN

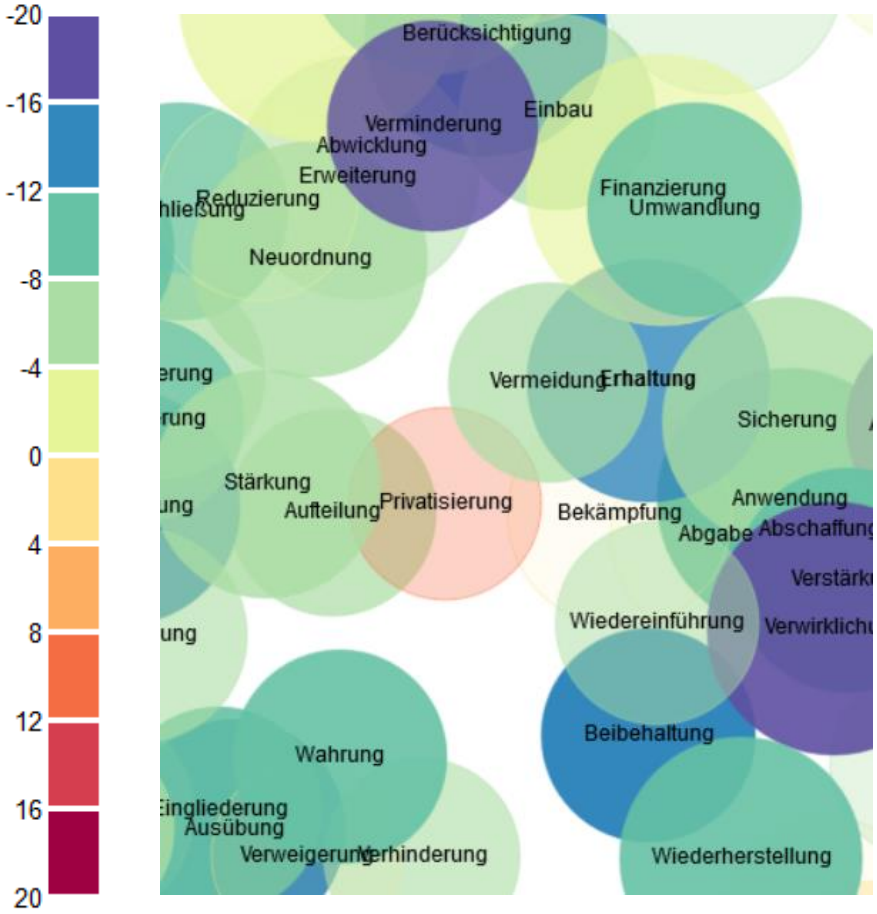
- Wort-Kookurrenz-Vektoren
  - $p(w_{-1}|w)$ (oder $PMI(w_{-1}, w)$)
  - Anzahl der Dimensionen = Vokabulargröße (* Kontextgröße)
- Worteinbettungen
  - Strukturiertes Skipgram (Wang2Vec [4])
  - Lerne Abbildung von Wort *w* auf Links/Rechts-Kontext $(w_{-2}, w_{-1,} w_{1,} w_{2,})$ über Zwischenebene mit wenigen Dimensionen (100-200)
- Diachrone Worteinbettungen [7]
  - Beginne mit randomisiertem Neuronalen Netz
  - Initialisiere Neuronales Netz zum Zeitpunkt $t + 1$ mit NN zum ZP t



VISUALISIERUNG VON WORTEINBETTUNGEN IN ZWEI DIMENSIONEN

- T-Distributed Stochastic Neighbor Embedding (T-SNE) [8]
  - Abbildung von *n* Dimensionen auf 2 Dimensionen
  - Gegeben: Wahrscheinlichkeit von Wortvektoren $x_i \; und \; x_j$: $p_{ij} = \frac{e^{-\left\|x_i - x_j\right\|^2 / 2\sigma^2}}{\sum_{k \neq l} e^{-\left\|x_k - x_l\right\|^2 / 2\sigma^2}}$
  - Finde: Wortkoordinaten $y_i \; und \; y_j$, mit: $q_{ij} = \frac{\left(1 + \left\|y_i - y_j\right\|^2\right)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}$
  - Sodass die KL-Divergenz zwischen *P* und *Q* minimiert wird:
    $KL(P||Q) = \sum_{i,j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right)$
- Erhält *lokale* Struktur, aber nicht die globale Struktur
  - Nahe x_i, x_j sollten nahe y_i, y_j entsprechen
  - Größere Distanzen (kleine p_{ij}) werden nicht sauber abgebildet.
- Daher: Globale Positionierung hat keine Interpretation

BEISPIEL: NOMINALISIERUNG MIT „-UNG“ VISUELLE KORRELATION



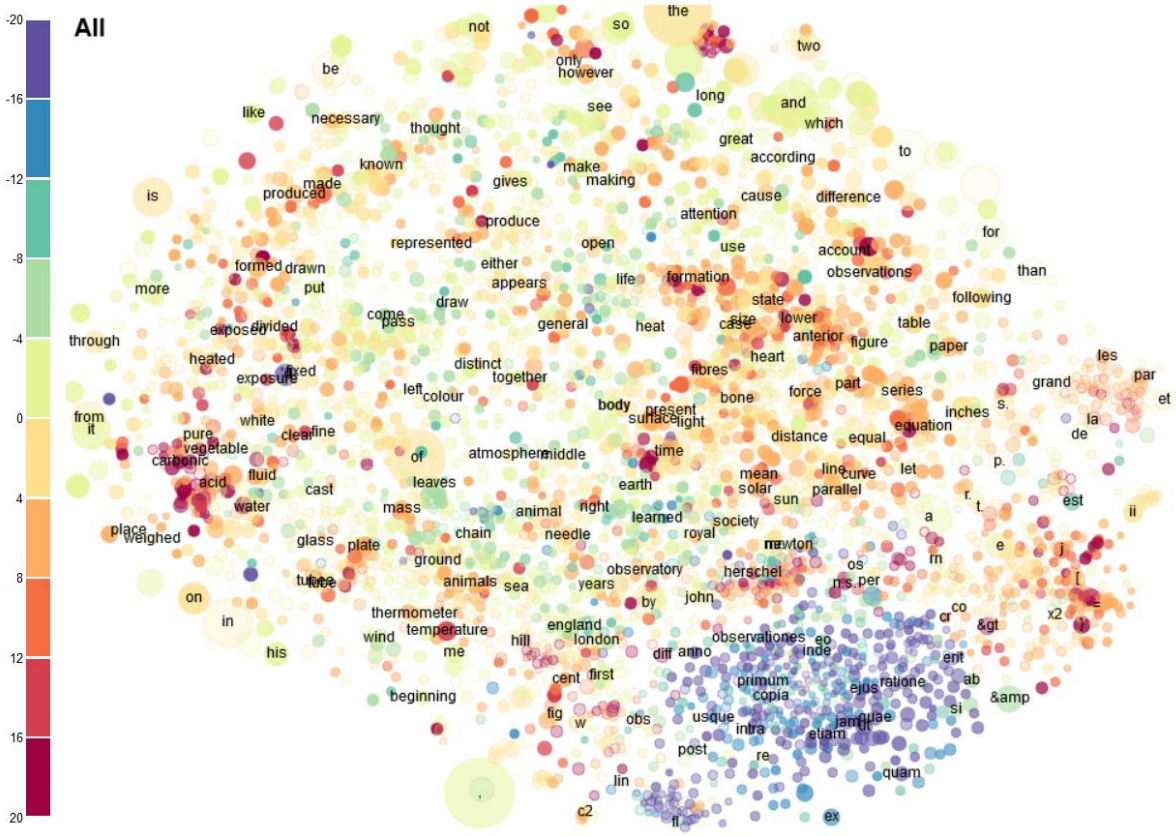
1955

2015

KORPORA: ÜBERSICHT

	Royal Society	Spiegel/Zeit	DeReKo Presse
Zeitspanne	1665-1869	1953-2015	2000-2015
Intervalle	10 Jahre	5 Jahre	1 Jahr
Token	35 Mio	570 Mio	18825 Mio
Visualisierte Typen	18700	25000	25000
R_1^2	31.1%	47.6%	44.4%
R_2^2	46.2%	66.6%	59.2%
Median Steigung	0.093	-0.014	-0.023
Dimensionen	100	200	200
Steigungskorrelation der nächsten Nachb.	0.77	0.43	0.42
http://corpora.ids-mannheim.de/openlab/diaviz/	royalsociety.html	zeitspiegel.html	dereko.html

VISUALISIERUNG ÜBERSICHT ROYAL SOCIETY KORPUS



KORRELATION ZWISCHEN FREQUENZÄNDERUNG UND GEBRAUCHSÄHNLICHKEIT (RSK)

- Koeffizienten der Frequenzänderung

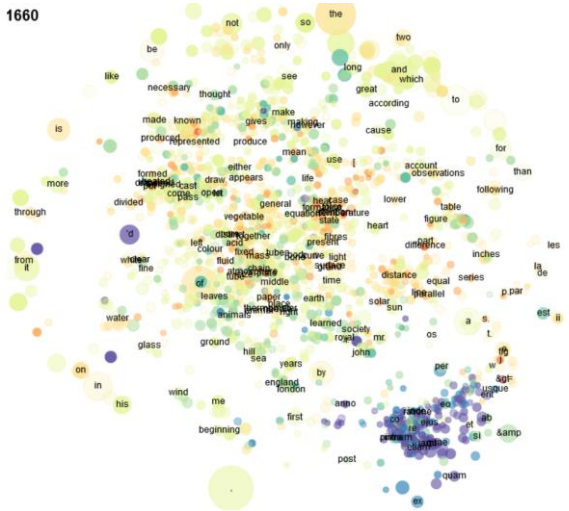
$$\ln\left(\frac{p(t)}{1-p(t)}\right) = k + s * t + c * t^2$$

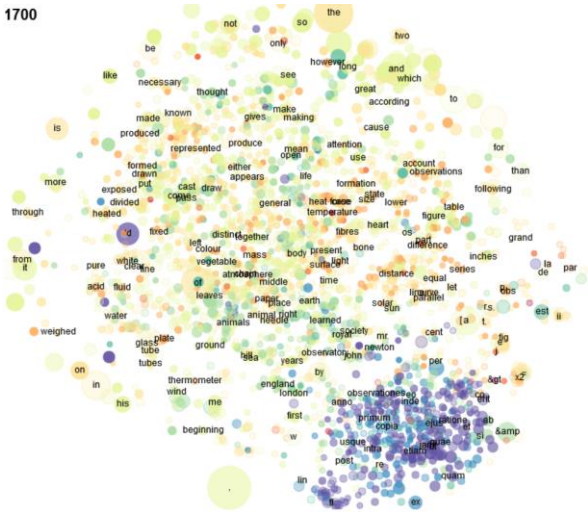
k … Intercept: Start
s … Steigung: Änderungsrate
c … Krümmung: Änderung der Steigung

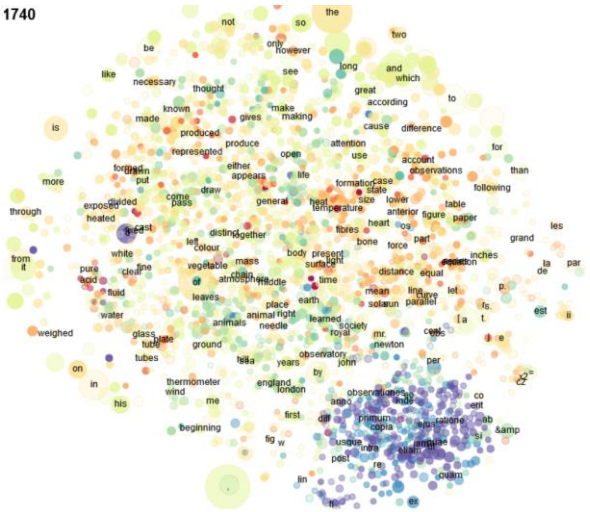
- Spearman-Rank-Korrelation ρ
  - Am stärksten zwischen den Steigungen der nächsten Nachbarn (*NN*)
  - Krümmung stärker korreliert als Intercept
  - Korrelation sinkt mit steigender Distanz zwischen NN (1,2,3).

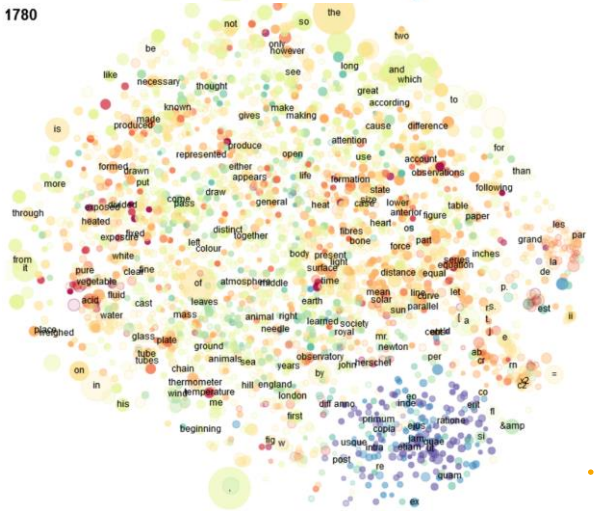
NN	*k*	*s*	*c*	
1	0.53	0.77	0.63	RSK
2	0.49	0.74	0.59	
3	0.46	0.73	0.56	
1	0.33	0.43	0.40	Spiegel/Zeit
2	0.27	0.39	0.33	
3	0.25	0.34	0.27	
1	0.33	0.42	0.48	DeReKo
2	0.26	0.36	0.41	
3	0.23	0.32	0.38	

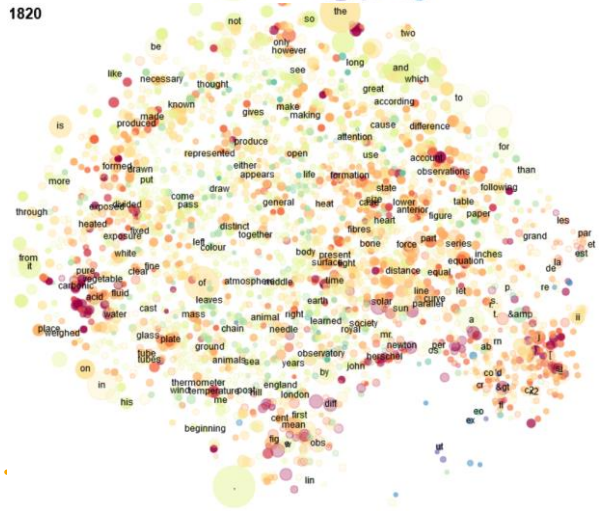
DIVERSIFIKATION DES VOKABULARS IM ROYAL SOCIETY KORPUS

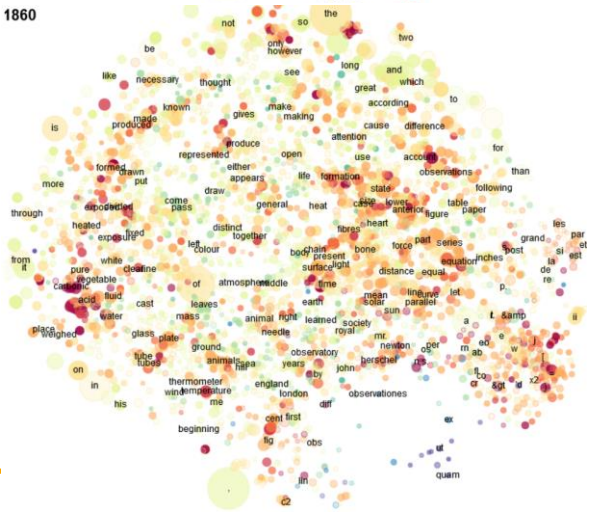




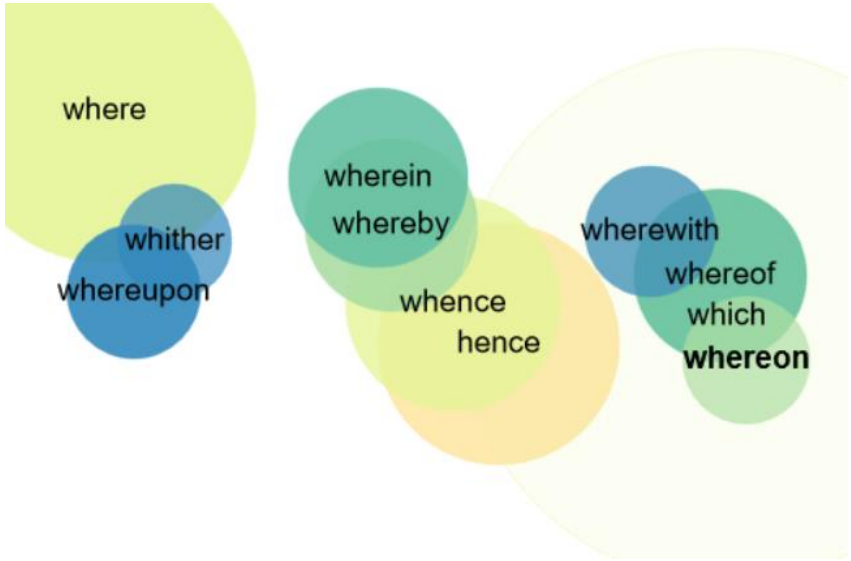


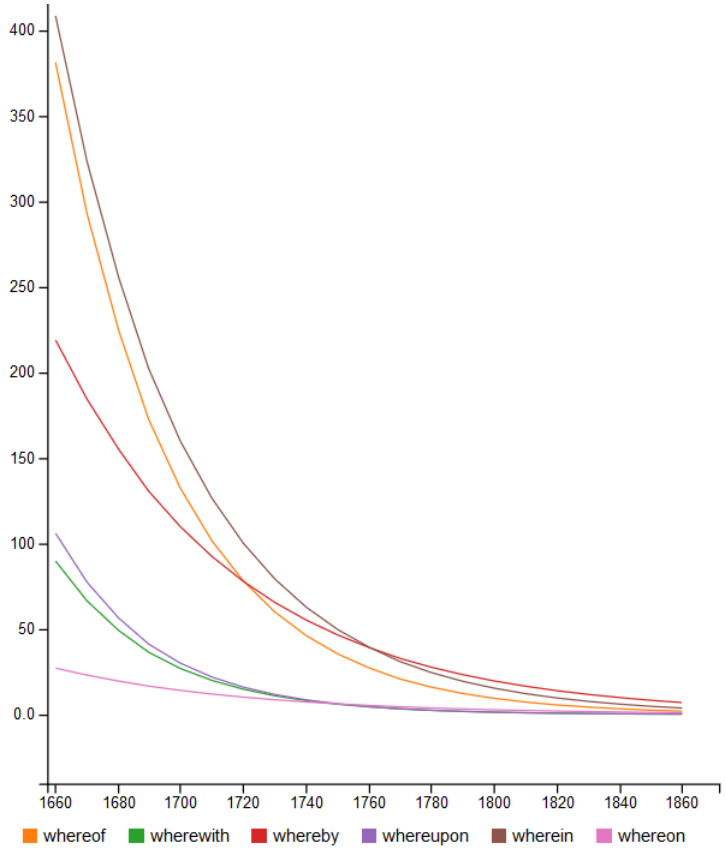






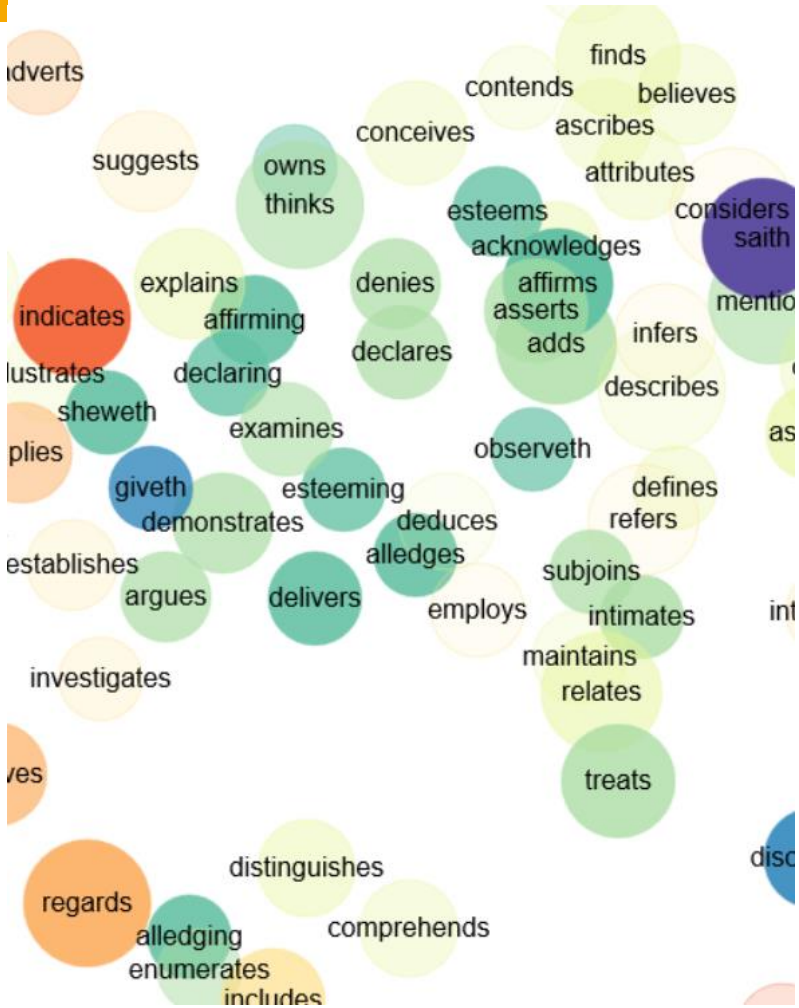
GRAMMATIK
WH-ADVERBIEN SINKEN



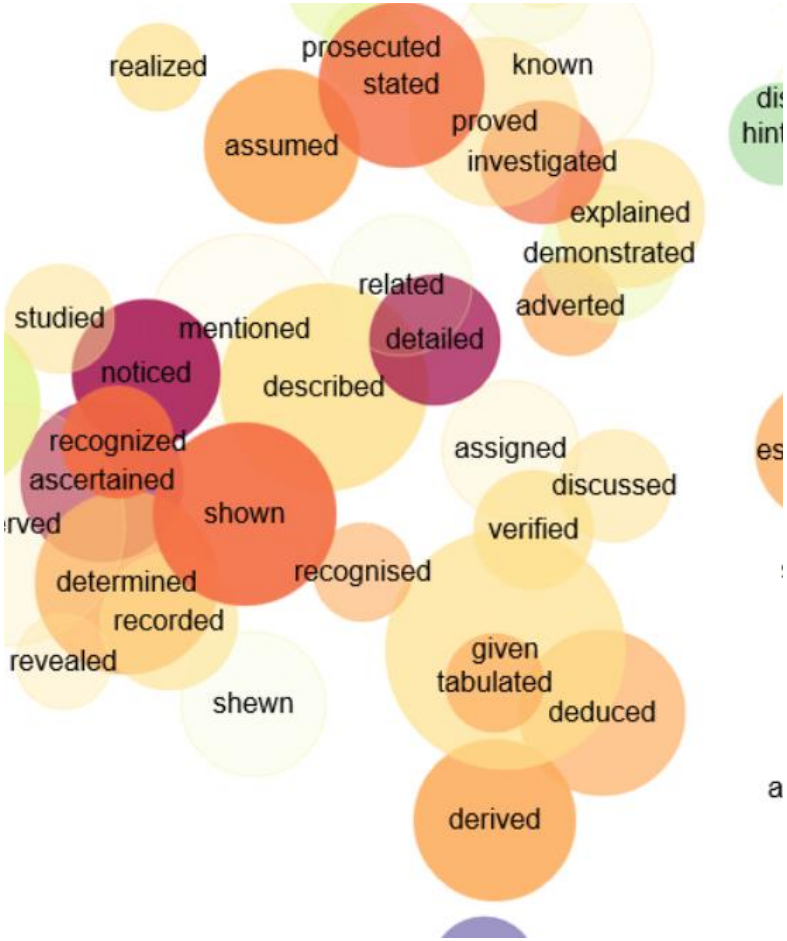


GRAMMATIK: KOMMUNIKATIONS/MENTALE VERBEN

PRÄSENS SINKT



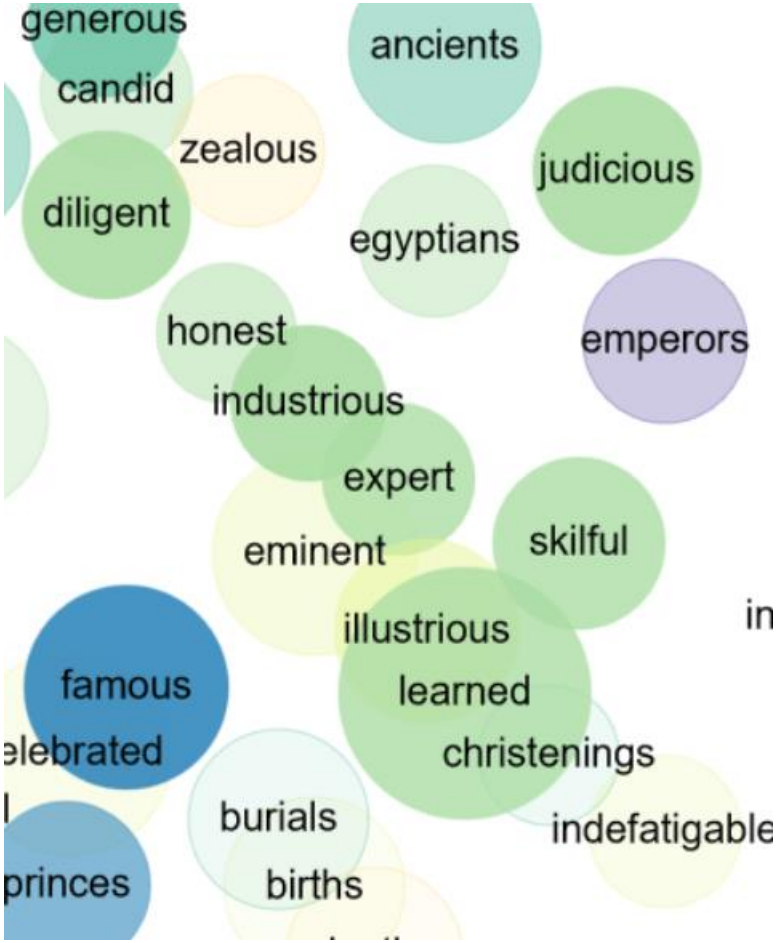
PASSIV/PERFEKT STEIGT

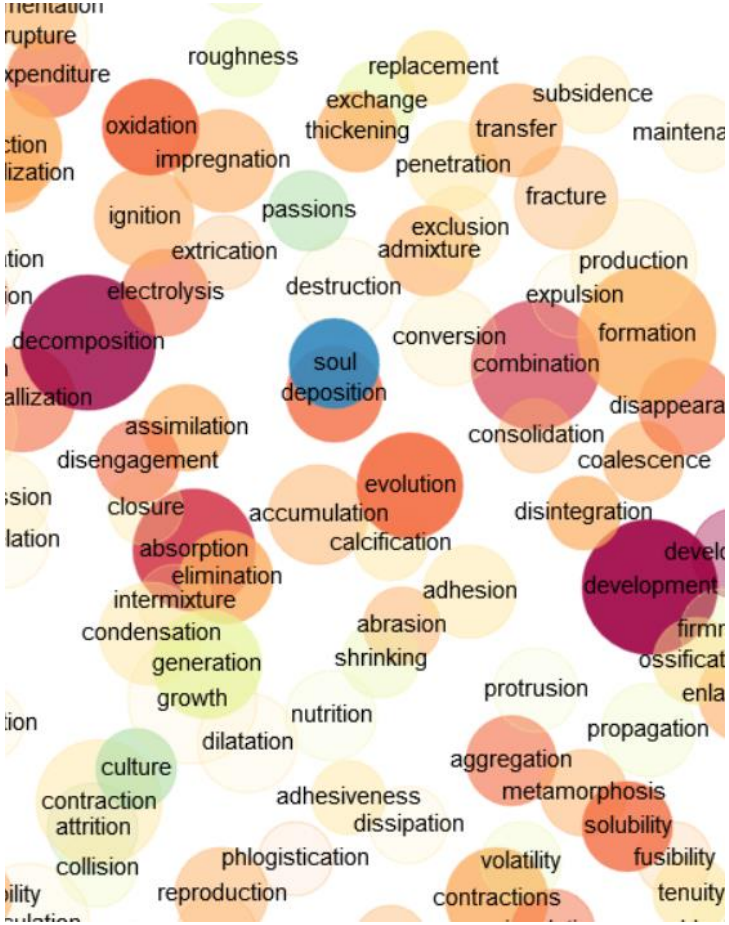


GRAMMATIK/STIL

PERSON-ADJEKTIVE SINKEN

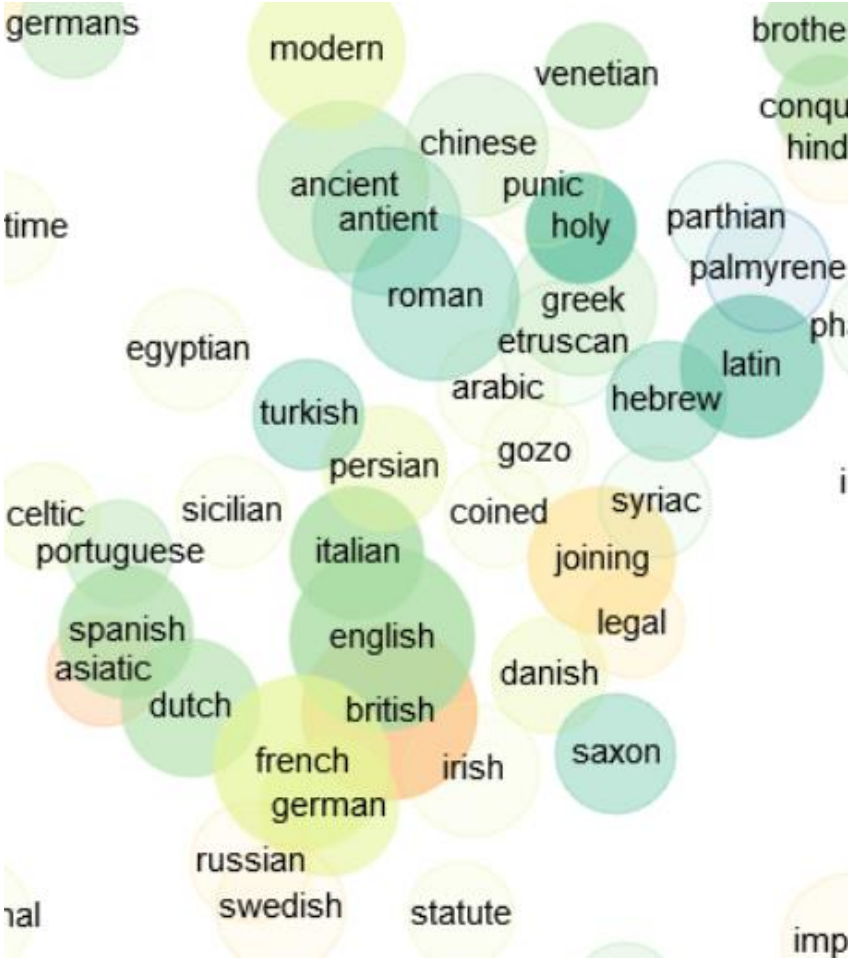
PROZESS-NOMINA STEIGEN



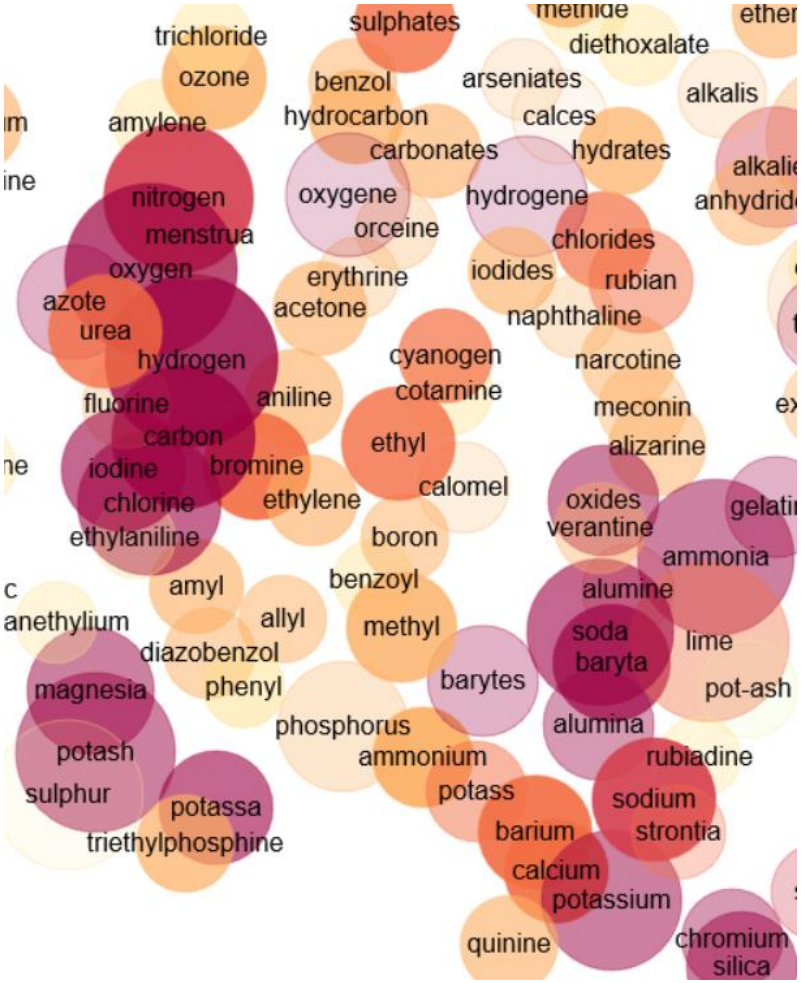


THEMA

NATIONALITÄT SINKT

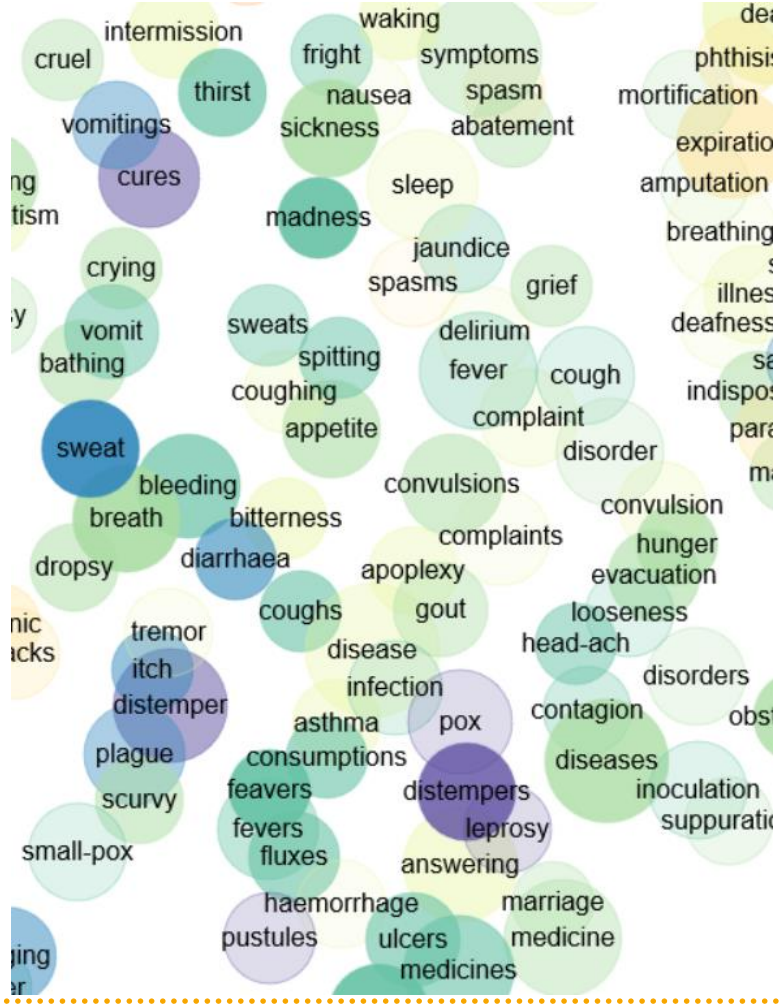


CHEMIE STEIGT

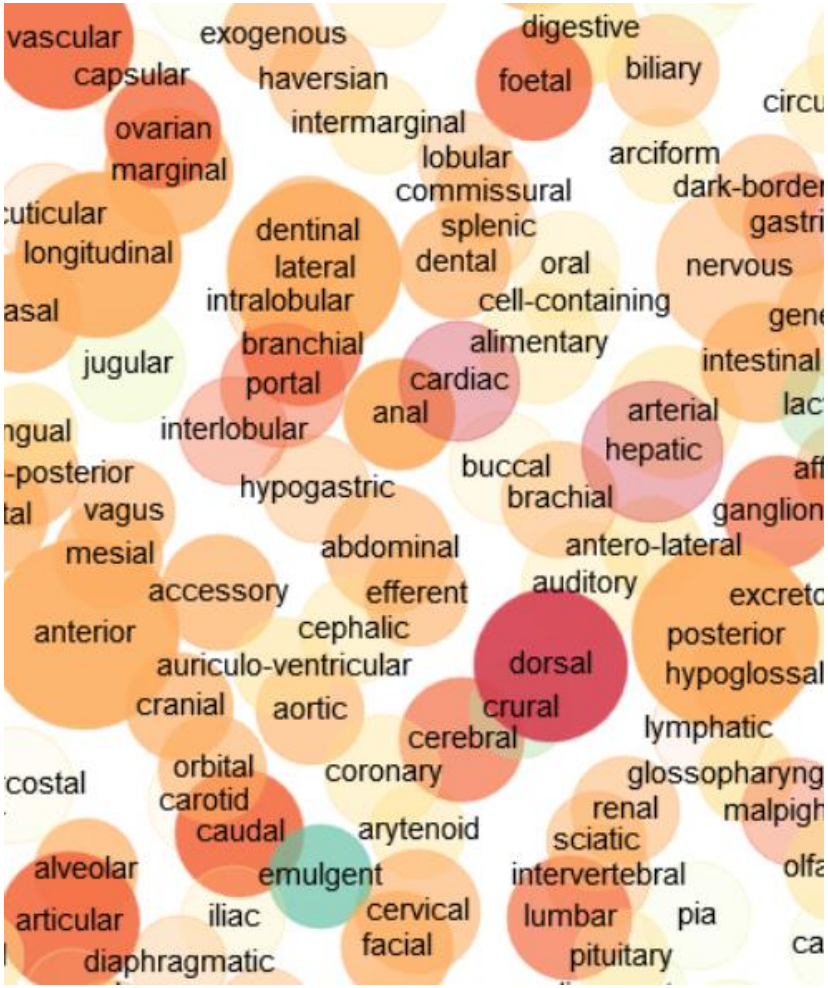


THEMA: MEDIZIN

SYMPTOME SINKEN



ANATOMIE STEIGT



NÄCHSTE NACHBARN: GEWINNER UND VERLIERER IM ROYAL SOCIETY KORPUS

Sinkt	Steigt
bigness	size
splendor	brilliancy
curious	interesting
loadstone	magnet
impertinent	unnecessary
plentiful	abundant
remembrance	recollection
hindred/hindered	prevented
contrived	constructed
truths	facts

NÄCHSTE NACHBARN. GEWINNER UND VERLIERER IM SPIEGEL/ZEIT KORPUS

Sinkt	Steigt
Carters	Obamas
DM	Euro
Brandt	Scharping
Herberger	Klinsmann
Rhodesien	Simbabwe
Industrialisierung	Globalisierung
Argwohn	Misstrauen
Erkenntnis	Gewissheit
fraglich	ungewiss
Grundbesitz	Immobilien

NÄCHSTE NACHBARN: GEWINNER UND VERLIERER IN DEREKO PRESSE

Sinkt	Steigt
Stoiber	Guttenberg
Tschernobyl	Fukushima
Windows	Android
Blair	Cameron
Scharping	Bahr
Kindergeld	Betreuungsgeld
PCs	Smartphones
Handys	Smartphones
Neuverschuldung	Schuldenbremse
Prospekte	Flyer
Selbstmord	Suizid

ZUSAMMENFASSUNG UND VERWANDTE ARBEITEN

- Paradigmatische Frequenzänderung
  - Paradigmatisch verwandte Wörter steigen oder fallen gemeinsam
  - Kurzfristig: Thema, Langfristig: Grammatikalische Präferenzen
- Lehrer 1985 [6]: Paralleler Wortgebrauchswandel
  - Paradigmatisch verwandte Wörter ändern/erweitern ihre Bedeutung gemeinsam
  - Beispiel: Ape/Baboon/Gorilla (abfällig)
  - Aber: Cf. Xu/Kemp 2015 [9]
- Kroch 1989 [5]: Hypothese der konstanten (gleichen) Änderungsrate (Steigung)
  - Sprachwandel erfolgt mit gleicher Steigung unabhängig vom Gebrauchskontext
  - Beispiel: Periphrastic Do
- Dubossarsky et al. 2015 [1]: „Marginale“ Wörter ändern häufiger ihre Bedeutung
  - Korrelation zwischen Distanz zwischen Wort und seinem Bedeutungszentrum und Bedeutungswandel
- Hamilton et al. 2015 [3]: Zwei „Gesetze“ von Bedeutungsänderung
  - Bedeutungsänderung vs. Frequenz vs. Polysemie

REFERENZEN

(1) Dubossarsky, H., Y. Tsvetkov, C. Dyer & E. Grossman (2015). A bottom up approach to category mapping and meaning change. In: Word Structure and Word Usage. Proceedings of the NetWordS Final Conference, 66-70. Pisa.

(2) Fankhauser, P., M. Kupietz (2017). Visualizing Language Change in a Corpus of Contemporary German. Corpus Linguistics Conference 2017.

(3) Hamilton, W., J. Leskovec & D. Jurawsky (2015). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. ACL 2016

(4) Kim, Y., Y. Chiu, K. Hanaki, D. Hedge & S. Petrov (2014). Temporal Analysis of Language through Neural Language Models. ACL 2014 Workshop on Language Technologies and Computational Social Science

(5) Kroch, A. (1989). Reflexes of grammar in patterns of language change. In Language Variation and Change 1 (3), 199-244.

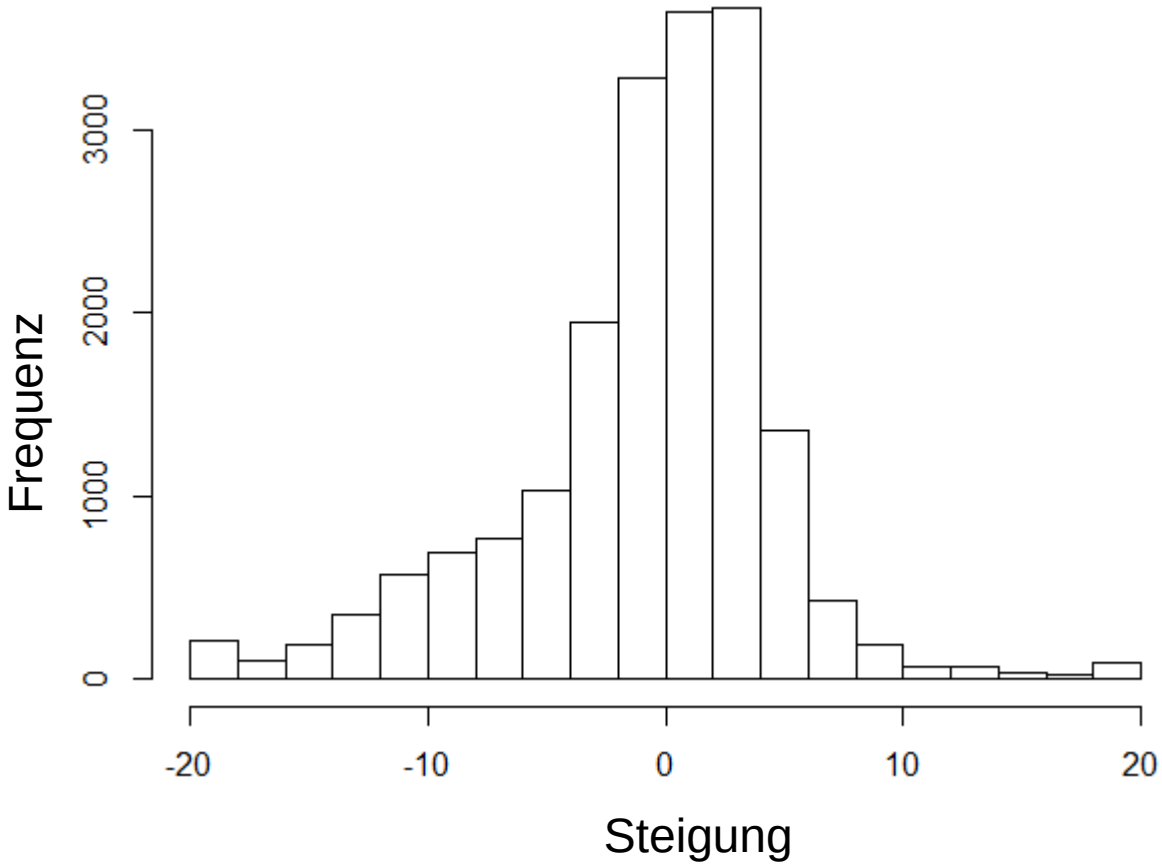
(6) Lehrer, A. (1985). The influence of semantic fields on semantic change. In J. Fisiak (Ed.), Historical semantics: Historical word formation (pp. 283-296). Berlin: Mouton de Gruyter.

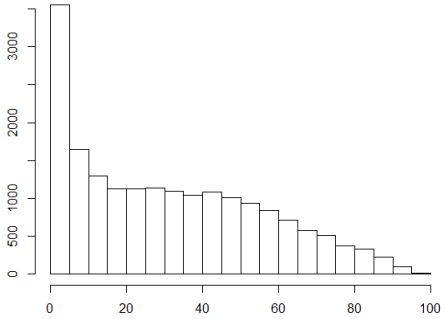
(7) Ling, W., C. Dyer, A. Black & I. Trancoso (2015). Two/Too Simple Adaptations of Word2Vec for Syntax Problems. Human Language Technologies (NAACL HLT 2015)

(8) Van der Maaten, L., G.E. Hinton (2008). Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9(Nov):2579-2605, 2008.

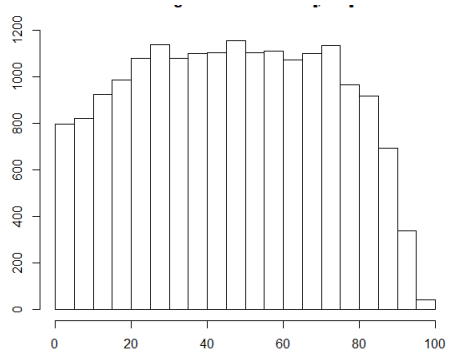
(9) Xu, Y. & C. Kemp (2015). A Computational Evaluation of Two Laws of Semantic Change. CogSci 2015.

RSK: VERTEILUNG VON STEIGUNG, PSEUDO-R2





R2(1): 31.1%



R2(2): 46.2%