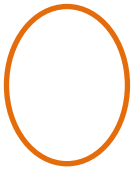
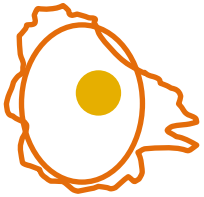
Korpuslinguistik

Was ist Bedeutung?

- Was bedeutet das Wort *Ei?*

Das Ei  ist im Kühlschrank.

Das Ei  brät in der Pfanne.

Was ist Bedeutung?

- Was bedeutet das Wort *nichts?*

Was ist Bedeutung?

- Was bedeutet das Wort *über*?

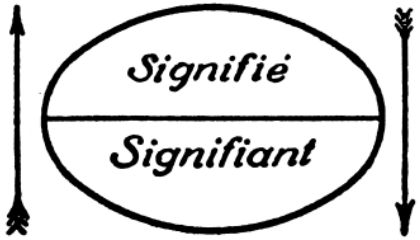
Das Bild hängt *über* dem Sessel.

Ich spreche *über* Peter.

Diese Präsentation hat *über* 100 Folien.

Was ist Bedeutung?

Zwei Ansätze – widersprüchlich oder komplementär?



- Saussure: sprachliches Zeichen hat eine Ausdrucks- und eine Inhaltsseite
- → repräsentationistische Bedeutungs-auffassung: sprachliche Zeichen stehen für außersprachliche Einheiten
- → "mentale" Bedeutungsauffassung: sprachliche Ausdrücke evozieren Vorstellungen
- Wittgenstein: "Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache."
- → gebrauchstheoretische Bedeutungsauffassung:

Frame-Semantik

- Frame-Semantik als "Verstehens-Semantik"
- Ein Frame ist eine komplexe Wissensstruktur, die durch sprachliche Zeichen evoziert wird
- keine Trennung von sprachlichem und enzyklopädischem Wissen

Charles Fillmore (1929-2014)

Was ist Bedeutung?

Das Weiße Haus steht kopf wegen Trumps neuer Frisur.

Was müssen wir wissen, um diesen Satz zu verstehen?

Was ist Bedeutung

Zwei Kannibalen essen einen Clown.

Sagt der eine: "Schmeckt irgendwie komisch."

Was müssen wir wissen, um diesen Witz zu verstehen?

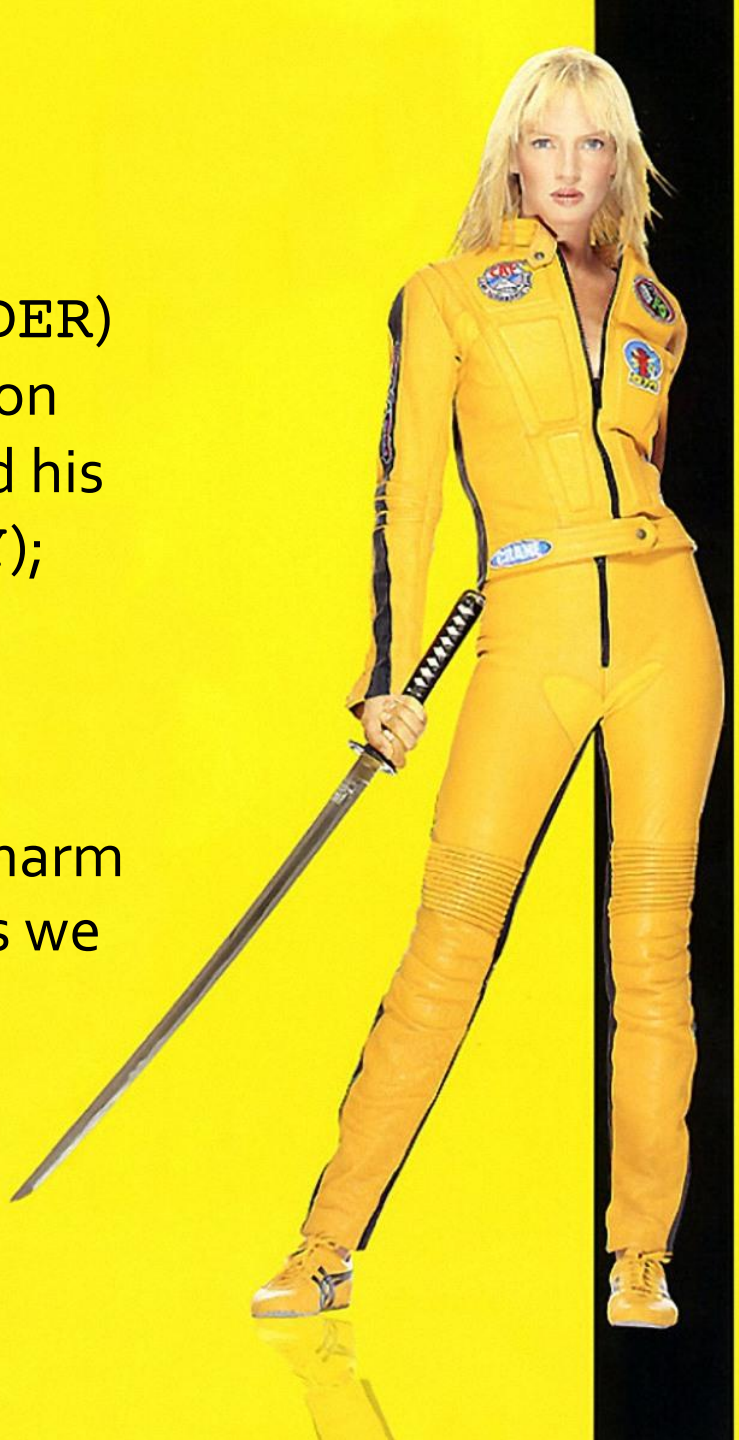
Frames

- standardisierte Formationen von Wissenselementen, die miteinander verknüpft sind
- werden mit "Leerstellen" und "Füllwerten" repräsentiert (*slots* und *fillers*)
- Beispiel: Frame RACHE beinhaltet verschiedene "Leerstellen" für GESCHÄDIGTE ENTITÄT, RÄCHER/IN, TÄTER/in
- zu unserem Frame-Wissen gehört, welche Entitäten die besagten Rollen füllen können und welche nicht
- z.B. kann eine Pflanze eher nicht die Rolle der RÄCHERIN ausfüllen

The Revenge frame

One person (we call him the OFFENDER) did something to harm another person (what he did we call the INJURY and his victim we call the INJURED_PARTY); reacting to that act, someone (the AVENGER, possibly the same individual as the INJURED_PARTY) acts so as to do harm to the OFFENDER, and what he does we call the PUNISHMENT.

(Boas 2017: 551)



Frame-Semantik

- befasst sich mit der Frage, wie sprachliche Ausdrücke Frames evozieren bzw. aktivieren
- Grundannahme: Alle gehaltvollen Wörter rufen eine Reihe von Frames auf, vor deren Hintergrund die Äußerung verstanden wird
- z.B. ruft das Verb *bezahlen* Frames auf wie VERKÄUFER, KÄUFER, GELD, GÜTER, PREIS…

Frame Element	Core Type
Buyer	Core
Circumstances	Extra-Thematic
Explanation	Extra-Thematic
Frequency	Extra-Thematic
Goods	Core
Manner	Peripheral
Means	Peripheral
Money	Core
Place	Peripheral
Purpose	Peripheral
Rate	Core
Seller	Core
Time	Peripheral
Unit	Peripheral

Bild: FrameNet, Lexical Unit *pay*, https://framenet2.icsi.berkeley.edu/fnReports/data/lu/lu3083.xml?mode=annotation

Semantik korpusbasiert untersuchen

- wichtige Erkenntnis der Frame-Semantik: Sprachliche Einheiten aktivieren umfangreiches (Hintergrund-)Wissen
- während sich z.B. (flexions-)morphologischer Wandel korpusbasiert einfach durch Beobachtung der Oberflächenstrukturen beschreiben lässt, ist die Untersuchung vonSemantik immer mit Interpretation verbunden.
- Wie kann man Semantik – und insbesondere semantischen Wandel – dennoch korpusbasiert untersuchen?

Semantischer Wandel: Wie *geil* ist das denn?

geil

geil Adj. ‘lüstern, geschlechtlich erregt’, in heutiger Jugendsprache ‘schön, großartig, toll’, ahd. geil ‘übermütig, überheblich, erhoben’ (8. Jh.), mhd. mnd. geil(e) ‘von wilder Kraft, mutwillig, üppig, lustig, begierig’, asächs. gēl ‘fröhlich, übermütig’, mnl. gheil, gheel ‘fröhlich, üppig, lüstern’, nl. geil ‘wollüstig’, aengl. gāl ‘lustig, lüstern, stolz’ (germ. *gaila- ‘fröhlich, lüstern’) und (mit Suffix erweitert) anord. geiligr ‘schön’ gehören vielleicht wie ablautendes mnl. ghīlen, nl. (älter) gijlen ‘gären, schäumen’, anord. gilker ‘Gärbottich’ mit lit. gailùs ‘scharf, beißend, bitter, kläglich’ und aslaw. ʒělo, russ. zeló (зело) ‘sehr’ zu ie. *ghoilos ‘aufschäumend, heftig, übermütig, ausgelassen, lustig’. (https://www.dwds.de/wb/geil)

geil

geil Adj. ‘lüstern, geschlechtlich erregt’, in heutiger Jugendsprache ‘schön, großartig, toll’, ahd. geil ‘übermütig, überheblich, erhoben’ (8. Jh.), mhd. mnd. geil(e) ‘von wilder Kraft, mutwillig, üppig, lustig, begierig’, asächs. gēl ‘fröhlich, übermütig’, mnl. gheil, gheel ‘fröhlich, üppig, lüstern’, nl. geil ‘wollüstig, geschlechtlich erregt’, engl. gaily ‘fröhlich, lüstern’) und (mit Suffix erweitert) anord. geiligr ‘schön’ gehören vielleicht wie ablautendes mnl. ghīlen, nl. (älter) gijlen ‘gären, schäumen’, anord. gilker ‘Gärbottich’ mit lit. gailùs ‘scharf, beißend, bitter, kläglich’ und aslaw. zělo, russ. zeló (зело) ‘sehr’ zu ie. *ghoilos ‘aufschäumend, heftig, übermütig, ausgelassen, lustig’. (https://www.dwds.de/wb/geil)

Woher weiß man das?!?

geil

- Doh waſ er fro vñ **geil** (Frauenfelder Flore, um 1220, REM)
- Also findest das in der geschrifft offt / der geist der hoffart / deß zorns / der geylheit / oder verbunsts / für die hoffertig / zornig / **geyl** vnnd verbünstig anfächtung gesetzt wirt. (Bullinger, Haußbuoch, 1558, DTA)
- Sie sindt das **allergeyleste** vnnd vnkeuscheste Volck so man in gantz Orient findet/ (Beatus, Amphitheatrvm Naturae, 1614, DTA)

Semantik korpusbasiert untersuchen

- Semantik ist nicht direkt beobachtbar
- korpusbasierte Untersuchungen zu Semantik und semantischem Wandel verlassen sich daher meist auf manuelle Annotation
- Vorteil: jeder Beleg wird im Kontext gesichtet → Aufdecken von Bedeutungsnuancen und Polysemien
- Nachteil: sehr zeitaufwändig, bei historischen Daten u.U. zu stark von gegenwartssprachl. Perspektive geprägt.

Semantik nach Firth

John Rupert Firth
(1890-1960):

You shall know a word by the company it keeps.

Explorative Methoden

- Unterscheidung zwischen **hypothesentestenden** und **explorativen** Methoden
- Zur Erinnerung: Hypothesentesten
  - Ich formuliere eine **Nullhypothese**…
  - … und versuche sie zu **falsifizieren**
- Explorativer Ansatz:
  - Keine explizite Hypothese
  - stattdessen: Ich versuche, **aus den Daten selbst** Muster zu erkennen

Distributionale Semantik

- geht davon aus, dass der **Kontext** wichtige Rückschlüsse auf semantische Eigenschaften sprachlicher Einheiten (meist: Wörter) zulässt
- zugrundeliegende Beobachtung: Wörter, die in ähnlichen Kontexten auftreten, haben meist auch ähnliche Bedeutungen.
- Herangehensweise: sog. *semantic vector space models*

Bag-of-words vector space model

Das mit der erfolgreichen Eroberung und "Zivilisierung" des "wilden Westens" einhergehende Bild des aufrechten
er erzählt von der Eroberung und Zivilisierung eines halben Kontinents .
als Mittel zur Eroberung und » Zivilisierung « fremden Bodens zu führen .
Es geht um Eroberung , vielleicht auch Zivilisierung .

Dabei ist es doch die eigene Kultur und Zivilisation , die sich als zu schwach erweist
Kultur ist Seele , ist Zeit ; Zivilisation ist Geist , ist Raum .
Was wir menschliche Kultur und menschliche Zivilisation nennen , sind Manifestationen solch einer Welt .
Die europäische Zivilisation hat den Endsieg über alle anderen Kulturen

erfolgreich
Eroberung
Boden
wild
Kontinent
Westen
einhergehen
Bild
Seele
Kultur
Manifestation
menschlich
europäisch
erweisen
Raum
Zeit
Welt

Wie weihnachtlich ist Spekulatius?

- Datenbasis: DECOW14AX
- Suche nach 9 Lemmata:
  - *Banane*
  - *Bonbon*
  - *Christstollen*
  - *Keks*
  - *Lebkuchen*
  - *Marzipan*
  - *Schokoriegel*
  - *Spekulatius*
  - *Zimt*

Wie weihnachtlich ist Spekulatius?

- Datenbasis: DECOW14AX
- Suche nach 9 Lemmata:
  - *Banane*
  - *Bonbon*
  - ***Christstollen***
  - *Keks*
  - ***Lebkuchen***
  - *Marzipan*
  - *Schokoriegel*
  - ***Spekulatius***
  - *Zimt*

Wie weihnachtlich ist Spekulatius?

- insgesamt 61.115 Belege
- Methode: 5 Lemmas aus dem linken Kontext, 5 Kommas aus dem rechten Kontext (keine Satzzeichen) gehen in die Analyse ein

von Spekulatius im September , auch **Spekulatius** und Lebkuchen echt lecker sein
dies typisch Weihnachtsleckerei wie Lebkuchen **Spekulatius** oder (unknown) mögen ich so
kalt , es duften nach Lebkuchen **Spekulatius** wir schlendern über Weihnachtsmarkt und
d Stadt und da sein deutsch **Spekulatius** unter ander sehr belieben
...

Kookkurrenz-Matrix

- Matrix aus Kookkurrenzfrequenzen
- insg. 204.075 Types im linken und rechten Kontext der 9 Keywords

Kollokate

Keywords

Lemma	Banane	Bonbon	Christstollen	Keks	Lebkuchen	Marzipan	Schokoriegel	Spekulatius	Zimt
abgucken	0	0	0	0	0	0	1	0	0
abhaben	0	0	0	0	2	0	0	0	0
abhalten	1	0	0	1	1	0	0	0	0
abhanden	0	0	0	0	0	0	1	0	0
Abhandlung	1	0	0	0	0	0	0	0	0
abhängen	1	0	0	0	1	0	0	0	0
abhängig	4	0	0	3	0	1	0	0	1
Abhängigkeit	2	0	0	0	0	0	0	0	0
abheben	0	1	0	0	0	0	0	0	0
Abheften	0	0	0	1	0	0	0	0	0
abhelfen	0	1	0	0	0	0	0	0	0
abhetzen	0	1	0	0	0	0	0	0	0
Abhilfe	0	2	1	0	0	0	0	0	1
abholen	1	3	0	13	1	1	1	0	0

Von Kookkurrenzen zu Vector Spaces

- Drei Schritte (vgl. Levshina 2015):
  - Errechnen von *Pointwise Mutual Information*-Werten durch den Vergleich von beobachten und erwarteten Kookkurrenz-Frequenzen,
  - Errechnen des Ähnlichkeitswerts mit Hilfe der Kosinus-Ähnlichkeit,
  - explorative Analyse der Ähnlichkeitswerte.

Schritt 1: (Positive) Pointwise Mutual Information (PPMI)

- $PMI(x,y) = \log_2(\frac{p(x,y)}{p(x)p(y)}) = \log_2(\frac{O_{xy}}{E_{xy}})$
- z.B.: *Keks* und *abhängig*…

Lemma	Banane	Bonbon	Christstollen	Keks	Lebkuchen	Marzipan	Schokoriegel	Spekulatius	Zimt
abgucken	0	0	0	0	0	0	1	0	0
abhaben	0	0	0	0	2	0	0	0	0
abhalten	1	0	0	1	1	0	0	0	0
abhanden	0	0	0	0	0	0	1	0	0
Abhandlung	1	0	0	0	0	0	0	0	0
abhängen	1	0	0	0	1	0	0	0	0
abhängig	4	0	0	3	0	1	0	0	1
Abhängigkeit	2	0	0	0	0	0	0	0	0
abheben	0	1	0	0	0	0	0	0	0
Abheften	0	0	0	1	0	0	0	0	0
abhelfen	0	1	0	0	0	0	0	0	0
abhetzen	0	1	0	0	0	0	0	0	0
Abhilfe	0	2	1	0	0	0	0	0	1
abholen	1	3	0	13	1	1	1	0	0

Schritt 1: Positive Pointwise Mutual Information (PPMI)

- *Keks* und *abhängig* treten dreimal zusammen auf
- Was wäre der **erwartete** Wert bei Zufallsverteilung?
- Antwort: 0.86558412
- (… warum, erfahren wir bei der Einführung in die Statistik Anfang nächsten Jahres!)

Schritt 1: Positive Pointwise Mutual Information

- **Positive** Pointwise Mutual Information: Wo immer PMI < 0, wird der Wert auf 0 gesetzt
- (Grund: bessere Ergebnisse, vgl. Bullinaria & Levy 2007)

Schritt 1: Positive Pointwise Mutual Information

- Was sagt uns der PPMI-Wert?
- Beispiel *Keks* und *Banane:* erwartet: 0.87, beobachtet: 3
- Formel also: PPMI = log2(3 / 0.87) = 1.79
- Je höher der beobachtete Wert im Vergleich zum erwarteten, desto höher PPMI
- Ist der beobachtete Wert kleiner als der erwartete, so ist PMI < 0, also PPMI = 0.

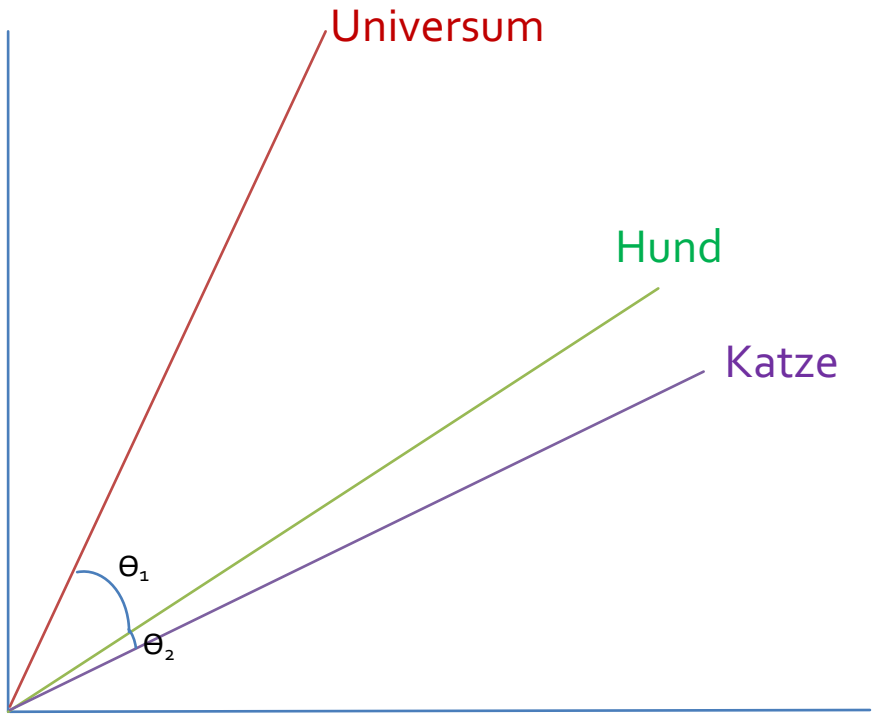
Schritt 2: Kosinus-Ähnlichkeit

- Wir haben nun eine Reihe von **Vektoren** mit PPMI-Werten

quadratisch	anfertigen	Podcast	nahend	Gesine	Paniermehl	mißlingen
1.397237	0.000000	1.397237	0.000000	0.000000	0.000000	0.000000

- hier: winziger Ausschnitt aus dem Vektor für *Banane*
- Die Vektoren werden quasi in Winkel überführt

Schritt 2: Kosinus-Ähnlichkeit



Fiktive Distributionsvektoren nach Levshina (2015)

Schritt 2: Kosinus-Ähnlichkeit

$$\cos(\Theta) = \frac{\sum_{i=1}^{n} a_i \cdot b_i}{\sqrt{\sum_{i=1}^{n} (a_i)^2} \cdot \sqrt{\sum_{i=1}^{n} (b_i)^2}}$$

- Wert reicht von -1 (genau entgegengerichtet) bis 1 (genau gleichgerichtet)

Schritt 3: Explorative Analyse der Ähnlichkeitsmaße

- Manuelle Inspektion der Resultate:

```
> round(christmas.cos,2)
             Banane Bonbon Christstollen Keks Lebkuchen Marzipan Schokoriegel Spekulatius Zimt
Banane         1.00   0.05          0.03 0.05      0.04     0.04         0.06        0.04 0.05
Bonbon         0.05   1.00          0.04 0.06      0.05     0.06         0.07        0.04 0.04
Christstollen  0.03   0.04          1.00 0.04      0.12     0.07         0.05        0.09 0.03
Keks           0.05   0.06          0.04 1.00      0.06     0.05         0.07        0.06 0.03
Lebkuchen      0.04   0.05          0.12 0.06      1.00     0.08         0.06        0.12 0.05
Marzipan       0.04   0.06          0.07 0.05      0.08     1.00         0.05        0.05 0.07
Schokoriegel   0.06   0.07          0.05 0.07      0.06     0.05         1.00        0.05 0.03
Spekulatius    0.04   0.04          0.09 0.06      0.12     0.05         0.05        1.00 0.04
Zimt           0.05   0.04          0.03 0.03      0.05     0.07         0.03        0.04 1.00
```

Schritt 3: Explorative Analyse der Ähnlichkeitsmaße

- Ähnlichkeitswerte können in **Distanzen** überführt werden.
  - Einfachste Möglichkeit: 1 – Ähnlichkeitswert
  - zwecks Normalisierung jedoch besser geeignet:

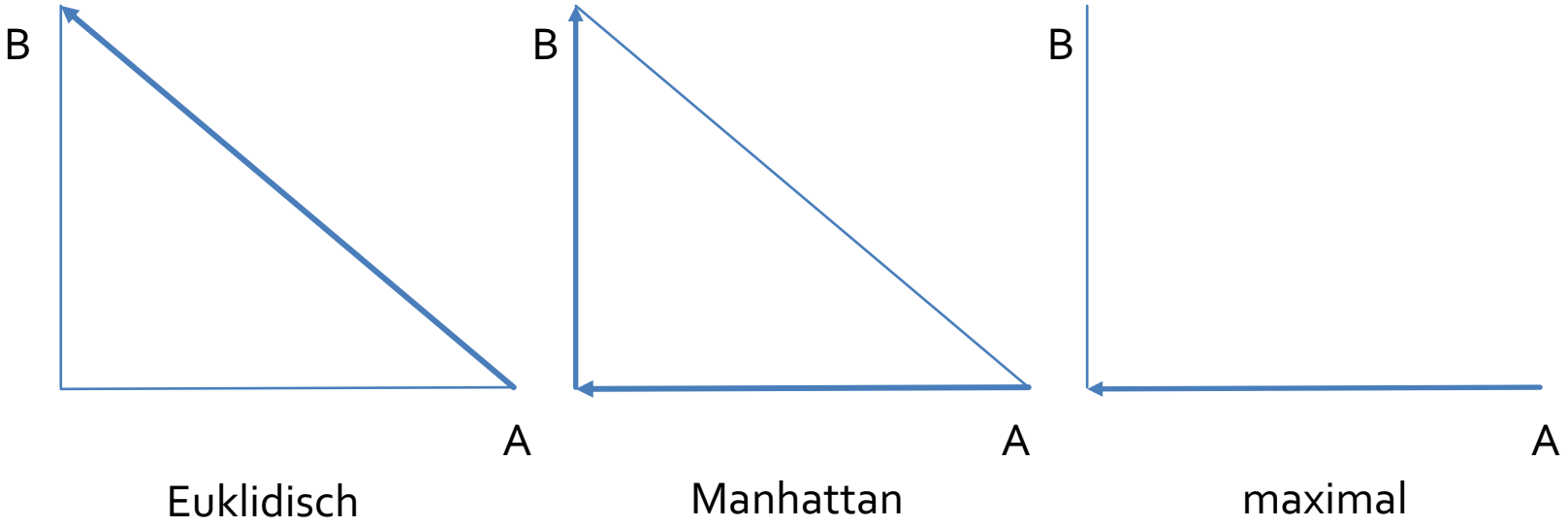
  1 –(aktueller Ähnlichkeitswert / maximaler Ähnlichkeitswert < 1)

Schritt 3: Explorative Analyse der Ähnlichkeitsmaße

- Wir wollen die Wörter nun in **Clustern** anordnen.
- Wie können wir die Cluster aus den Daten identifizieren?
- Lösung: "Partitioning around medoids"

Schritt 3: Explorative Analyse der Ähnlichkeitsmaße

- PAM teilt die Elemente um Medoide herum auf, wobei ein Medoid ein zentrales Exemplar ist, von dem aus die Distanz zu allen anderen Mitgliedern des Clusters minimal ist
- Distanz kann auf verschiedene Weise bestimmt werden



Schritt 3: Explorative Analyse der Ähnlichkeitsmaße

- Anzahl der Cluster lässt sich manuell bestimmen:
  - *Banane*
  - *Bonbon*
  - *Christstollen*
  - *Keks*
  - *Lebkuchen*
  - *Marzipan*
  - *Schokoriegel*
  - *Spekulatius*
  - *Zimt*

Schritt 3: Explorative Analyse der Ähnlichkeitsmaße

- Anzahl der Cluster lässt sich manuell bestimmen:
  - *Banane*
  - *Bonbon*
  - *Christstollen*
  - *Keks*
  - *Lebkuchen*
  - *Marzipan*
  - *Schokoriegel*
  - *Spekulatius*
  - *Zimt*

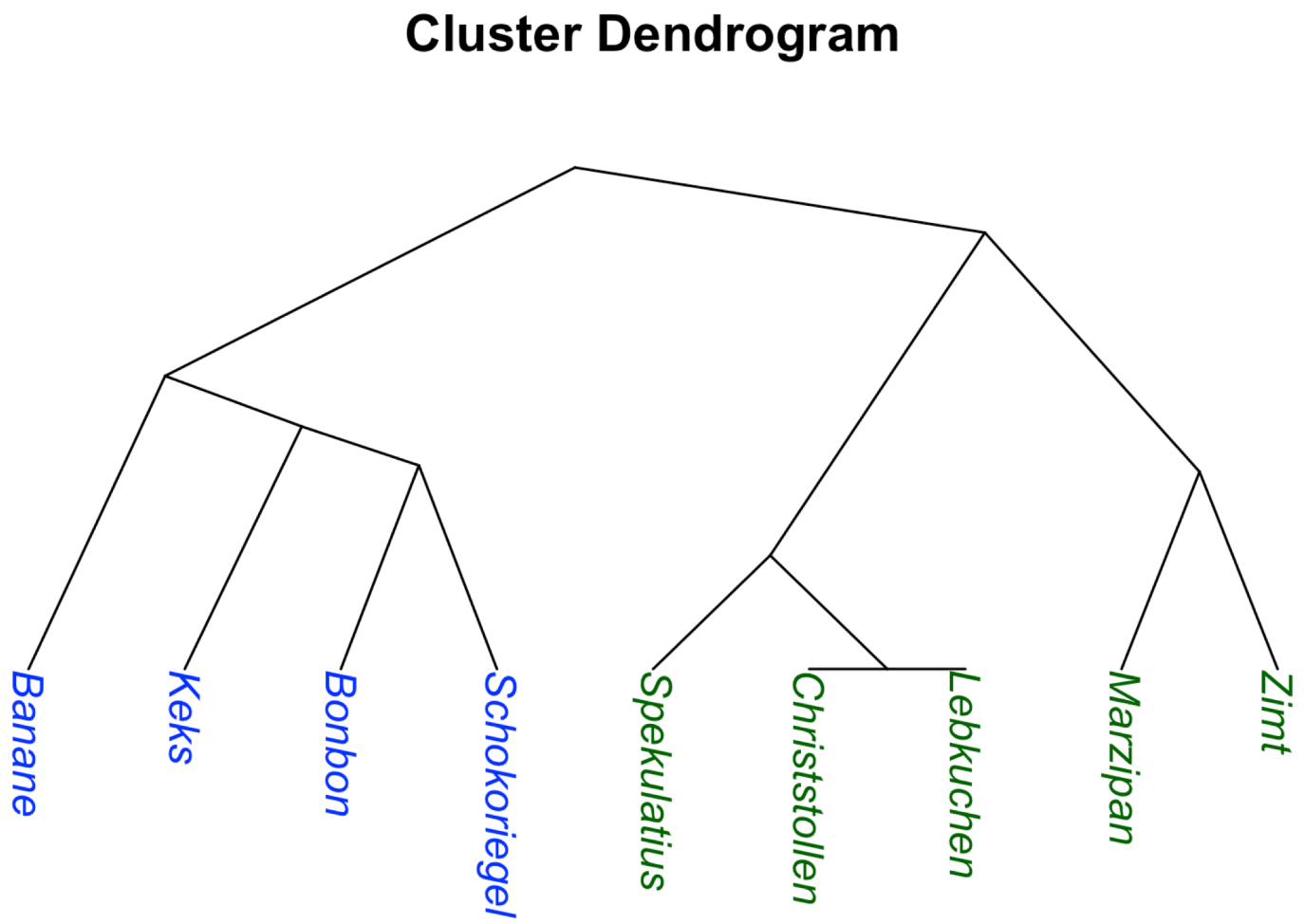
Schritt 3: Explorative Analyse der Ähnlichkeitsmaße

- Anzahl der Cluster lässt sich manuell bestimmen:
  - *Banane*
  - *Bonbon*
  - *Christstollen*
  - *Keks*
  - *Lebkuchen*
  - *Marzipan*
  - *Schokoriegel*
  - *Spekulatius*
  - *Zimt*

Schritt 3: Explorative Analyse der Ähnlichkeitsmaße

- Welche Anzahl an Clustern ist die beste?
- "Average Silhouette Width" gibt Auskuft
- zeigt, wie wohlgeformt die Cluster sind
- Wert zwischen 0 und 1
- 0: keine Clusterstruktur in den Daten
- 1: klare Struktur, vollständige Trennung zwischen den Clustern.
- bester Wert hier: zwei Cluster.

Ergebnis: Hierarchical Clustering

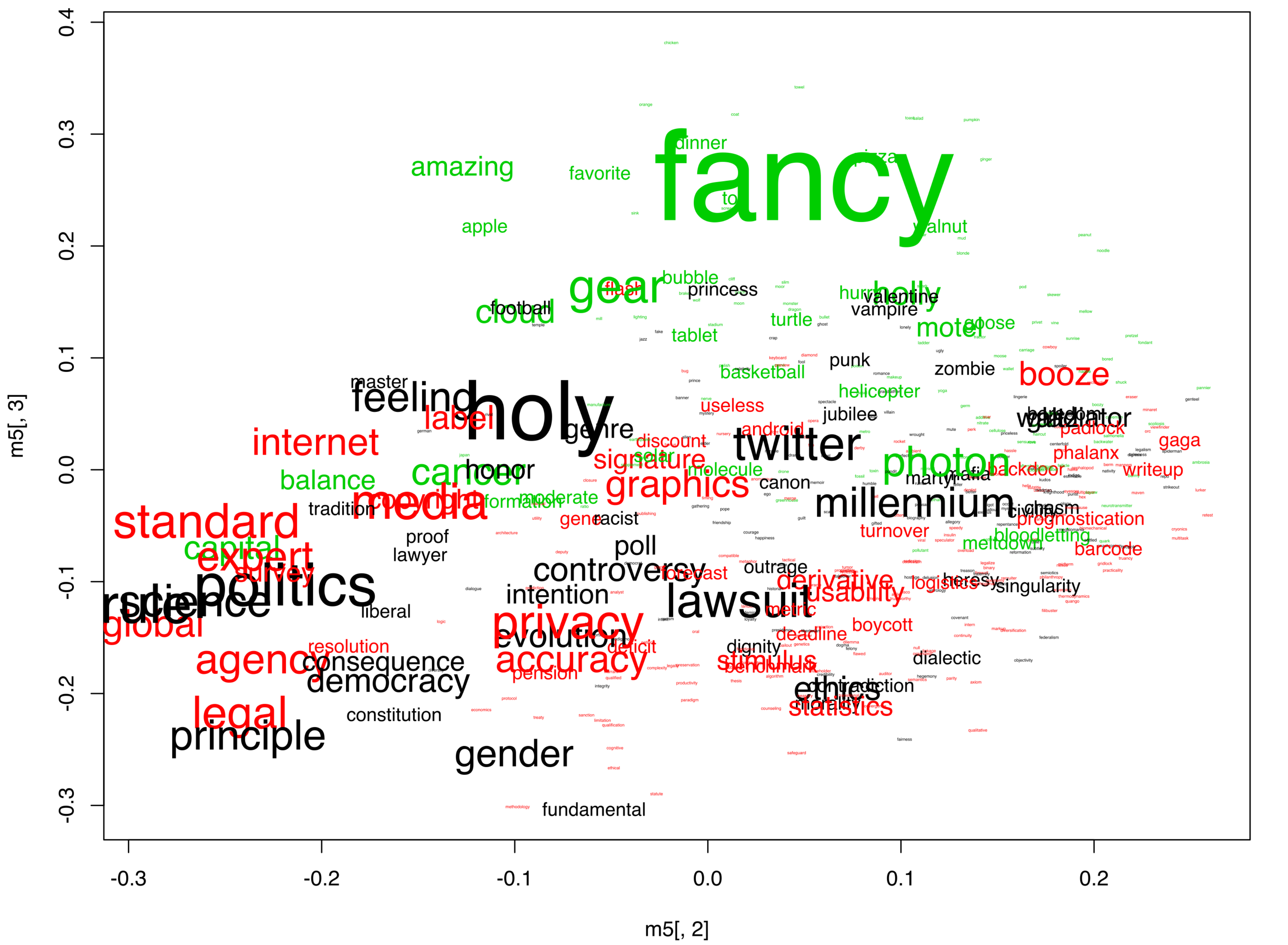


Fazit

- Hohe Übereinstimmung zwischen "intuitivem" und datenbasiertem Clustering
- Daher interessante Methode gerade für ältere Sprachstufen, bei der "intuitive" Urteile über die Semantik von Wörtern und Konstruktionen nur sehr bedingt möglich sind.

Vector-space model für shm-reduplication

- zur Erinnerung: *legal, schmegal* u.ä.
- Datengrundlage: ENCOW16-Daten
  - Suche nach allen Lemmata, die in der Konstruktion vorkommen
  - Extraktion von KWICs aus dem gesamten Korpus (5 Wörter links, 5 Wörter rechts)
  - Erstellen einer Ähnlichkeitsmatrix, Errechnen der Kosinus-Distanz
  - Clustering mit *partitioning around medoids* (PAM)



m5[, 3]
m5[, 2]
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.3
-0.2
-0.1
0.0
0.1
0.2
fancy
dinner
amazing
favorite
apple
walnut
gear
bubble
princess
football
cloud
turtle
tablet
valentine
vampire
motel
goose
basketball
punk
zombie
booze
helicopter
useless
jubilee
holy
feeling
master
label
genre
internet
android
twitter
discount
signature
graphics
photon
balance
honor
molecule
canon
millennium
media
tradition
moderate
racist
standard
proof
lawyer
poll
capital
expert
politics
controversy
forecast
outrage
turnover
meltdown
bloodletting
prognostication
barcode
phalanx
writeup
gaga
padlock
backdoor
mafia
liberal
intention
lawsuit
derivative
usability
logistics
heresy
singularity
privacy
evolution
global
metric
boycott
deadline
resolution
deficit
dignity
dialectic
agency
consequence
accuracy
democracy
pension
stimulus
benchmark
contradiction
ethics
statistics
constitution
legal
principle
gender
fundamental

Übung: REM und DTA/DWDS

- Einfache Aufgabe: Wir sehen uns die Kontexte von *Kopf* im REM und DTA an