ASYMMETRIE IN MEHRWORTAUSDRÜCKEN:
„krasser Unterschied“ vs *„himmelweiter Unterschied“*

Yana Strakatova (Tübingen)

Kookkurrenzen

1. Kollokat + Basis:

großer Unterschied, fürstliches Gehalt, blinder Passagier, schwarzes Schaf

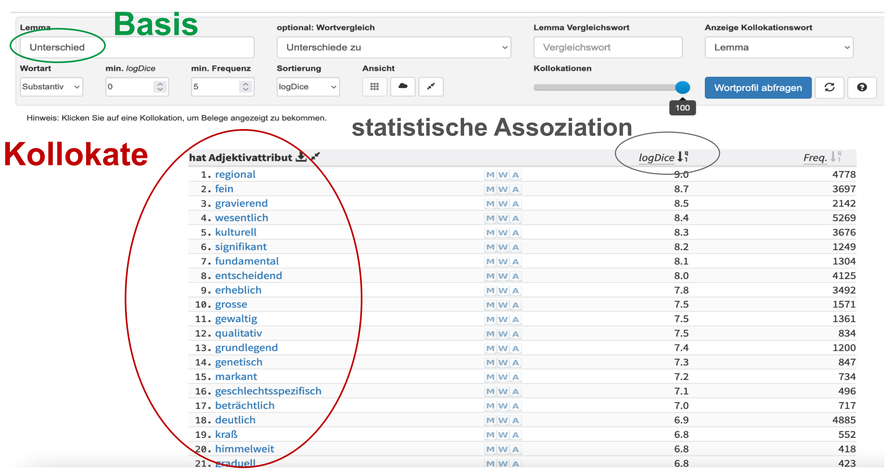
2. Statistische Assoziation:

Berechnet basierend auf Korpusfrequenzen („contingency table“ [1])

Statistischer Assoziationswert **logDice =** $14 + log_2 \frac{2a}{(a+c)+(a+b)}$ [2]

- theoretische Obergrenze = 14, die meisten Werte < 10
- unabhängig von Korpusgröße
- negative Werte = keine statistische Signifikanz

DWDS-Wortprofil [3]



Kollokat (logDice)

wesentlich (8.4)

beträchtlich (7.0)

deutlich (6.9)

krass (6.8)

himmelweit (6.8)

Contingency table

	Nomen	¬Nomen	∑
Adjektiv	a	b	a+b
¬Adjektiv	c	d	c+d
∑	a+c	b+d	a+b+c+d=n

a – Frequenz des Bigrams; **a+c** – Frequenz des Nomens;
a+b – Frequenz des Adjektives

Asymmetrie

krass: Unterschied, Gegensatz, Kontrast, Gegenteil, Ungleichheit, Diskrepanz, Abweichung, usw.

himmelweit: Unterschied

Asymmetrischer statistischer Assoziationswert: **ΔP** [4]
Werte zwischen -1 und 1

left-predictive: ΔP1

$$\Delta P_1 = \Delta P(w_1|w_2) = \frac{a}{a+c} - \frac{b}{b+d}$$

frühe Morgenstunde

right-predictive: ΔP2

$$\Delta P_2 = \Delta P(w_2|w_1) = \frac{a}{a+b} - \frac{c}{c+d}$$

himmelweiter Unterschied

Wie häufig ist solche Asymmetrie? Welche Arten von Mehrwortausdrücken sind asymmetrisch? Kann man asymmetrische Werte in das DWDS-Wortprofil einbauen?

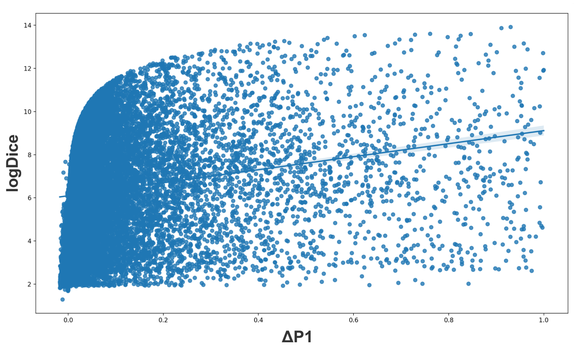
Korpusstudie

Daten

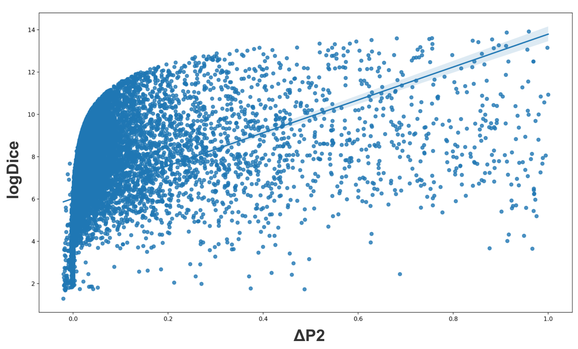
- Korpus (Wikipedia 2017, 2018 [5]; decow16ax [6])
- Frequenz > 400
- 36,945 AN-Kombinationen
- Statistische Werte: ΔP1, ΔP2, logDice
- Manuelle Annotation als „Kollokation“, „Begriff“, „Eigenname“, „Freie Phrase“, „Idiom“, „Klischee“

Korrelation mit logDice

alle AN-Kombinationen

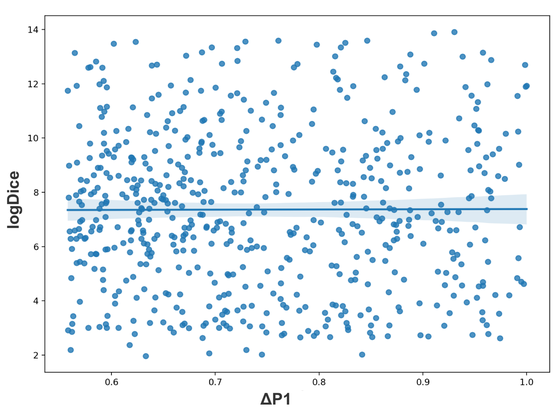


Pearson=0.184,
Spearman=0.376

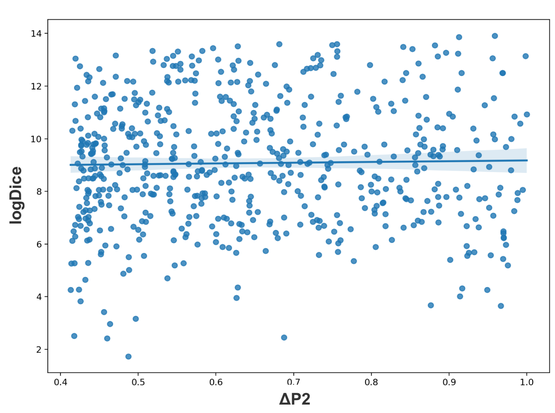


Pearson=0.373,
Spearman=0.877

top-600 AN-Kombinationen



Pearson=0.003,
Spearman=-0.01



Pearson=0.021,
Spearman=0.023

600-best right-predictive (ΔP2)

Label	Anzahl	Beispiele
Begriff	256	*rheumatoide Arthritis (0.97), glykämischer Index (0.76), festverzinsliche Wertpapiere (0.64), molare Masse (0.54)*
Kollokation	182	*geraume Zeit (0.98), triftiger Grund (0.89), brenzlige Situation (0.84), faustdicke Überraschung (0.70), klirrende Kälte (0.57)*
Eigenname	87	*Timmendorfer Strand (0.93), (der) Gestiefelte Kater (0.90), Brühlsche Terrasse (0.73), (Das) Wohltemperierte Klavier (0.46)*
Idiom	15	*gordischer Knoten (0.89), zweischneidiges Schwert (0.61), (am) seidenen Faden (hängen) (0.41)*

600-best left-predictive (ΔP1)

Label	Anzahl	Beispiele
Eigenname	240	*Olympische Sommerspiele (0.96), (der) Letzte Mohikaner (0.87), Hohe Tatra (0.62), Vereinigtes Königreich (0.56)*
Begriff	208	*vorläufige Vollstreckbarkeit (0.95), chronische Polyarthritis (0.83), absoluter Nullpunkt (0.68), Neue Sachlichkeit (0.65)*
Freie Phrase	65	*passender Staubsaugerbeutel (0.95), langer Blütenstiel (0.60), früher Sonntagmorgen (0.56)*
Kollokation	40	*letzte Ruhestätte (0.91), üble Nachrede (0.90), reine Formsache (0.89), gesunder Menschenverstand (0.87), stehende Ovation (0.79), schwere Kopfverletzung (0.61)*
Idiom	28	*innerer Schweinehund (0.85), eierlegende Wollmilchsau (0.85), ewige Jagdgründe (0.69), fliegende Untertasse (0.69)*
Klischee	6	*zu guter Letzt (0.91), gutes Gelingen (0.79), herzlichen Glückwunsch (0.78), gute Besserung (0.68)*
Kompositum	1	*rote Beete (0.63)*

References: **[1]** Stefan Evert. 2004. The Statistics ofWord Cooccurrences: Word Pairs and Collocations. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. **[2]** Pavel Rychly. 2008. A Lexicographer-Friendly Association Score. In Sojka, Petr /Hor.k, Aleš (Hg.): Proceedings of the 2nd Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2008, pages 6–9, Brno. **[3]** *DWDS-Wortprofil für „Unterschied“, erstellt durch das Digitale Wörterbuch der deutschen Sprache, <https://www.dwds.de/wp/Unterschied>, abgerufen am 26.09.2023.* **[4]** Michelbacher, L., Evert S., Schütze, H., Asymmetry in corpus-derived and human word associations. Corpus Linguistics and Linguistic Theory, 2021, 7(2):245 – 276. **[5]** Daniel de Kok and Sebastian Pütz. 2019. Stylebook for the Tübingen Treebank of Dependency-parsed German (TüBa-D/DP). Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, Germany. **[6a]** Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pages 486–493, Istanbul, Turkey. European Language Resources Association (ELRA). **[6b]** Roland Schäfer. 2015. Processing and querying large web corpora with the COW14 architecture. In Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3), Lancaster. UCREL, IDS.