

Charakterisierung von Redensarten und Phraseologismen mittels statistischer Textanalyse

Programmierpraktikum
Institut für Computerlinguistik, Universität Zürich

Dieter Bachmann
dab@flaez.ch

März 2004

Inhaltsverzeichnis

1	das Problem	2
1.1	Redensarten	2
1.2	Problemstellung	3
2	Methode	4
2.1	Einschränkung des Problems	4
2.2	die Korpora	5
2.3	Lemmatisierung	6
2.4	Kollokation	7
2.5	semantische Felder	8
2.5.1	Schuhe, Teufel	10
3	Resultate	11
3.1	Zusammenfassung	13

1 das Problem

1.1 Redensarten

Die Erkennung der redensartlichen oder figurativen Verwendung eines Begriffs setzt Weltwissen voraus, und letztlich totales Textverständnis. Nicht nur das, es ist auch für den menschlichen Leser in vielen Fällen nicht abschliessend entscheidbar, ob eine bestimmte Wendung nun als figurativ zu betrachten ist oder nicht: gerade in diesem Bereich des spielerischen und kreativen Umgangs mit der Sprache werden Autoren oft gerade deshalb zu einer bestimmten Redensart greifen, weil sie im aktuellen Kontext ambig sein kann.

Dazu kommt, dass viele figurative Umschreibungen sich derart *abschleifen*, dass sie zum unmarkierten Begriff werden, und ihr ursprünglich redensartlicher Charakter nur noch als Etymologie gelten kann (z.B. das Verb *einleuchten* wird kaum mehr als figurativ wahrgenommen, während *heimleuchten* noch etwas drolliger wirkt und eher ein Bild heraufbeschwört).

Selbst wenn sich eine bestimmte redensartliche Wendung auf ein fixiertes Sprichwort bezieht, wird man kaum je den genauen Wortlaut dieses Sprichworts vorfinden, also etwa *“wer einem eine Grube gräbt, fällt selbst hinein”* wird kaum so in einem Zeitungsartikel stehen, sondern schon die Wörter *Grube* und *graben* zu einem Satz verbaut werden genug distinktiv sein, zu signalisieren, dass auf das Sprichwort verwiesen wird.

Je nachdem, wie mutig und wie stilsicher ein Journalist ist, können solche Anspielungen auch verunglücken, so dass die beabsichtigte Referenz unklar wird, oder man beim Lesen über Widersinn oder Absurditäten stolpert:

Nach der Wahrnehmung eines Diplomaten üben sich zusehends mehr Iraker im Halslockern, um diesen dann rechtzeitig wenden zu können. (NZZ von 28. 2. 2003)¹

Solche Verwendung irgendwie automatisch als redensartlich zu erkennen wird praktisch hoffnungslos sein, da das distinktive *Hals* gar nicht genannt wird (es sei denn, man wolle sich auf Komposita-Segmentierung einlassen, was die Gefahr einer Flut von *false positives* bürge).

Verschiedene Typen von Redensarten können daran unterschieden werden, wie *kompositionell* (*compositional*; [2]) sie sind, d.h. in welchem Mass sie sich mit ihrem nicht-metaphorischen Kontext verbinden können, ohne ihre Idiomatik einzubüssen ([1]; [4], zitiert nach [8]). Diese Kompositionalität / *compositionality* ist fließend, und kann je nach Autor mehr oder weniger strapaziert werden. Allgemein gilt, dass eine Redensart desto leichter erkennbar sein wird, je weniger kompositionell sie ist, insbesondere natürlich, wenn eines ihrer Glieder derart obsolet oder selten ist, dass es nicht mehr un-idiomatisch verwendet werden kann (*Argusaugen*, *Fettnäpfchen*, (?)*Wässerchen*)².

¹In diesem Beispiel zeigt das anaphorische Pronomen ins Leere, weil zwar vom *Halslockern*, nicht aber vom *Hals* selbst die Rede war. Ausserdem ist *Wendehals* redensartlich, aber *Halslockern* ist eine ad-hoc-Bildung und erscheint unelegant.

²Lexikostatistisch müsste man aufhören, von Redensarten zu sprechen, wenn ein Wort (wie *Fettnäpfchen*) nur noch in der neuen Verwendung vorkommt, und sich von seinem ‘Begleiter’ zu lösen beginnt (*treten* ist nicht mehr Voraussetzung zur Kennzeichnung des nicht-wörtlichen Gebrauchs.)

Der einfachste Weg, und für praktische Anwendungen der wohl einzig gangbare, ist, mit einer fixen Liste von Phraseologismen zu operieren, also etwa

{(Schuh, drücken), (Grube, graben), (Wässerchen, trüben), (Leber, kriechen), ...}.

Für die damit gefundenen Redensarts-Kandidaten muss dann jeweils aus dem Kontext eine Entscheidung gefällt werden, wie wahrscheinlich im konkreten Fall eine idiomatische Verwendung vorliegt.

Ein solches Vorgehen ist starr und wird der proteischen Natur gerade dieses Bereichs der Sprache nicht gerecht: Die Redensarten werden lexikalisiert, d.h. die Verantwortung einer abschliessenden Aufzählung liegt beim Verfasser des Lexikons und das Problem wird damit gar nicht Teil der eigentlichen Textanalyse.

1.2 Problemstellung

Ziel dieser Arbeit ist die Untersuchung der Verteilung von Phraseologismen in einem Beispieldatensatz. Es soll dabei rein statistisch vorgegangen werden, d. h. es wird auf die Vorgabe einer Liste von Redensarten verzichtet. Dass ein solches Vorgehen niemals praktische Anwendbarkeit erreichen kann, liegt auf der Hand. Vielmehr soll es darum gehen, Erkenntnisse über die Signatur der statistischen Verteilung redensartlicher Wendungen zu gewinnen, die in praktischen Anwendungen im Verbund mit optimierten Lexika und detaillierter grammatischer Analyse mithelfen können, den Status einer bestimmten Wendung zu bestimmen.

Dabei lasse ich mich von folgenden Grundideen leiten:

- Redensarten werden nach ihrer Prägung von ihrem Ursprungs-Kontext unabhängig weiter tradiert und haben oft ein beträchtliches Alter. Sie gehören gleichsam einer älteren Textschicht an, was sich in altertümlichem Vokabular und nicht selten in morphologischen Archaismen (Dativ-*e* etc.) niederschlägt.
- Redensarten kommen durch Kombination von zwei oder mehr Elementen zustande, die bei je isoliertem Auftreten keinen redensartlichen Charakter haben und oft zum unauffälligen Grundvokabular gehören.
- Redensarten werden figurativ/metaphorisch gebraucht, d. h., ihre Konstituenten gehören normalerweise *nicht* ins semantische Feld ihrer Umgebung im Text.
- Redensarten wiederholen sich und sind dem Leser bekannt.

Für jeden dieser Punkte lassen sich Gegenbeispiele anführen:

- Neuschöpfungen. Hier würde ich weniger von Redensarten als von Floskeln und *buzz words* sprechen. Man könnte argumentieren, dass der Übergang von einer Floskel zu einer Redensart ein allmählicher, organischer Prozess ist, dem man mit der Bezifferung einer "Phraseologie-Score" eher gerecht wird als mit einem Lexikon.
- Univerbierungen wie *Teufelskreis* erscheinen nur noch als eingliedrig und werden mit diesem Vorgehen nicht erfasst.

- Es gibt durchaus Redensarten, die immer noch wörtlich verwendet werden. Zum Beispiel “für das leibliche Wohl sorgen” hat eindeutig redensartlichen Charakter, wird aber wörtlich verwendet (im semantischen Feld Gastronomie etc.). Auch sonst kommt es oft vor, dass eine Redensart von semantisch verwandten Begriffen begleitet wird; entweder, weil sich gleich eine andere Redensart aus einem verwandten Gebiet an sie anschliesst, oder durch Triggern einer Redensart durch einen mit dem Bild inhaltlich verwandten Kontext; im folgenden Beispiel etwa stehen die beiden Ausdrücke mit *Teufel* in semantischem Zusammenhang (und allenfalls wurde das Lexem *Teufel* bereits attrahiert durch das vorangehende *katholisch*):

Das Vabanquespiel der existenziell gebeutelten Katholisch-Konservativen erinnert an den Teufel, der in der Not Fliegen frisst. Doch die Sozialdemokraten muss der Teufel selber reiten, [...] (NZZ vom 6. 12. 2003)

- Ein menschlicher Leser kann eine neue Redensart lernen, ohne dass sie deklariert wird (Er muss Redensartlichkeit inferieren, wenn der Text sonst keinen Sinn macht). Bei der bescheidenen Grösse meiner Korpora darf ich nicht damit rechnen, dass sich eine bestimmte Redensart wiederholt.

Ungeachtet dieser Einwände will ich versuchen, nach den genannten Kriterien die Redensartlichkeit von Ausdrücken in einem Beispielkorpus zu bewerten.

2 Methode

Zur Beurteilung von verschiedenen Vokabular-Schichten verwende ich neben dem Testkorpus einen Kontrollkorpus. Die Sprache des Kontrollkorpus ist wesentlich älter (klassische Literatur) als die des Testkorpus (Zeitungstext). Damit soll simuliert werden, dass ein Leser einen Zeitungstext versteht vor dem Hintergrund anderer Sprachregister (Literatur, Soziolekt usw.), die beim Auftreten von Redewendungen mit der aktuellen Textsorte interferieren.

2.1 Einschränkung des Problems

Um den Rechenaufwand in realistischen Grenzen zu bewahren, beschränke ich mich auf relativ kleine Korpora. Damit fällt eine wichtige Eigenschaft von Redensarten, ihr Wieder-Auftreten (Reokkurrenz), für uns weg (*sparse data*). Im Rahmen dieser Arbeit kann es nur darum gehen, aus der vollen Kombinatorik eine Auswahl von Redensarten-Kandidaten zu filtern (die dann anhand von grösseren Korpora auf Reokkurrenz getestet werden könnten). Für uns überflüssig wird damit auch eine Beurteilung der Syntax (z. B. ‘*Grube graben* + Dativ’), weil die Möglichkeit fehlt, das konsistente Auftreten einer bestimmten Syntax zu überprüfen.

Daneben wächst der Aufwand der Berechnung von Kollokationen exponentiell mit der Anzahl der beteiligten Komponenten. Deshalb soll hier davon ausgegangen werden, dass eine Redensart durch zwei Wörter determiniert ist, also im Stil der oben genannten $\{(Schuh, drücken), (Grube, graben), (Wässerchen, trüben), (Leber, kriechen)\}$. Redensarten wie “auf grossem Fuss(e) leben”, “in der Not frisst der Teufel Fliegen”, die aus mehr als zwei obligatorischen Gliedern bestehen, können in der Regel auch durch zwei ihrer Glieder erkannt werden

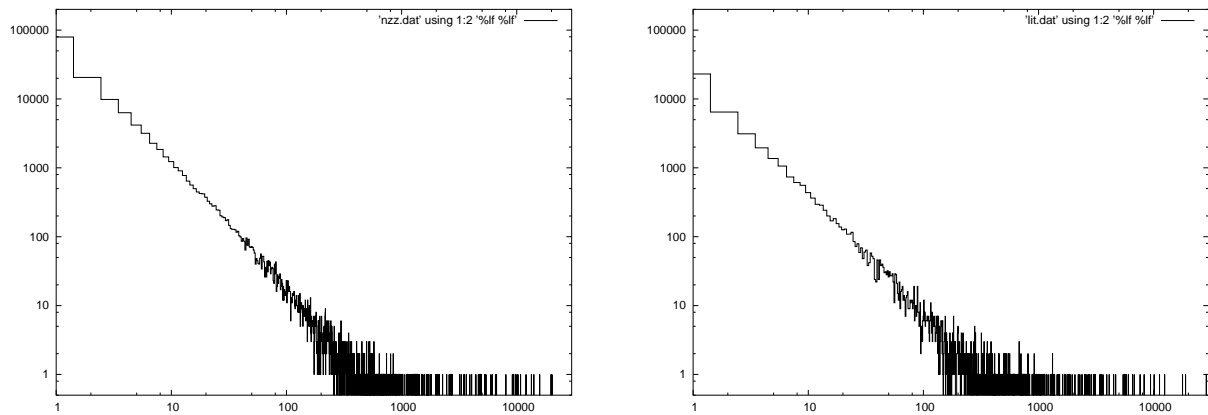


Abbildung 1: Verteilung der Worthäufigkeiten in den beiden Korpora (Zipf's law)

(*semantic priming*; [3]) und ihre Identifizierung wird somit nicht von vornherein ausgeschlossen, wird aber unwahrscheinlicher, weil das dritte Glied den bewerteten semantisch an den Text binden kann.

Weiter beschränke ich mich auf Redensarten, die aus einem Nomen und einem (Verb oder Adjektiv) bestehen (gemäss der Bestimmung durch den UIS-Parser); rein nominale Ausdrücke ($\{Kind, Kegel\}$, $\{Stock, Stein\}$) fallen damit also ausser Betracht; nicht weil sie mit der verwendeten Methode schlechter erfassen liessen, sie interessieren hier aber weniger, weil sie sich weniger in ein Satzgefüge einpassen und daher leichter durch einfache Kollokation erkannt werden als verbale Ausdrücke.

2.2 die Korpora

Als Testkorpus verwende ich den Text aller Ausgaben der NZZ von April 1994 (Kürzel `nzz`, Grösse 14M). Dieser Korpus besteht aus 1'495'825 Wörtern, davon 145'688 unterschiedliche (9.6%).

Als Kontrollkorpus dient mir eine Zusammenstellung klassischer deutscher Literatur³ (Kürzel `lit`, Grösse 5.5M). Dieser Korpus hat besteht aus 701'933 Wörtern, davon 44'802 unterschiedliche (6.3%).

Das deutlich grössere Vokabular des NZZ-Korpus dürfte auf die grosse Zahl von Eigennamen zurückgehen.

Die Schnittmenge des Vokabulars der beiden Korpora zählt 20'327 Einträge, d.h. nur 14% der Wörter in `nzz` erscheinen auch in `lit`. Abb. 2 zeigt die Verteilung der Frequenzverhältnisse dieser in beiden Korpora vorhandenen Wörter, gewichtet nach Korpusgrösse: Sind die Anzahl Vorkommnisse eines Wortes W in den Korpora beschrieben durch $\text{freq}_{\text{nzz}}(W)$ und $\text{freq}_{\text{lit}}(W)$, so zeichne ich

³eine Auswahl aus dem (wertvollen, aber leider benutzerunfreundlichen) "Projekt Gutenberg-DE" Korpus (`gutenberg2000.de`); sie umfasst die Autoren Buechner, Fontane, Goethe, Grimm, Heine, Kafka, Lessing, Schiller, Spyri, Storm und Wieland.

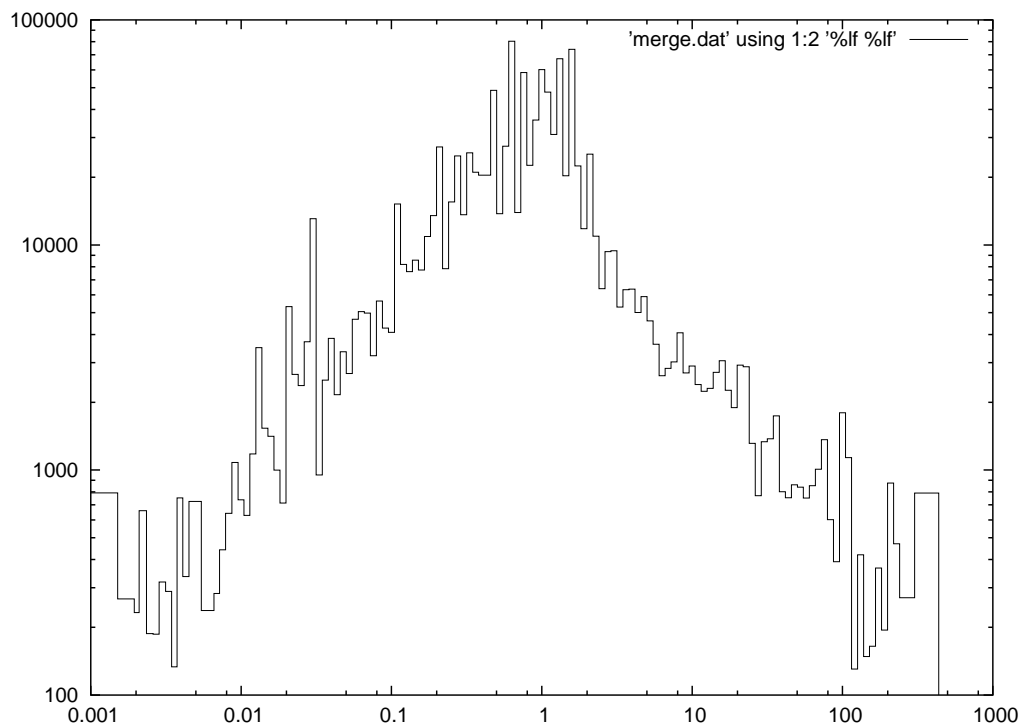


Abbildung 2: Frequenzverhältnisse (siehe Text)

$$\frac{\text{freq}_{\text{nzz}}(W) \cdot r}{\text{freq}_{\text{lit}}(W)} \text{ vs. } (\text{freq}_{\text{nzz}}(W) \cdot r + \text{freq}_{\text{lit}}(W))$$

mit $r = \frac{701933}{1495825} = 47\%$ dem Grössenverhältnis der Korpora. D. h., in der Grafik erscheinen links die Wörter, die häufiger in `lit`, rechts die Wörter, die häufiger in `nzz` auftreten. Man kann sehen, dass Wörter, die in einem der Korpora extrem viel häufiger auftreten etwa gleichmässig verteilt sind. Abb. 3 zeigt explizit eine Auswahl der extrem verteilten Wörter; man kann daran z. B. sehen, dass die 2. Person in der Zeitung viel seltener verwendet wird.

2.3 Lemmatisierung

Weil wir erwarten, dass die für die Redensart distinktiven Vokabeln im Text verbaut werden, möchten wir die Wörter in den Korpora auf ihre Stammformen reduzieren.

Mir stand als Werkzeug zur Lemmatisierung der UIS-Parser⁴ zur Verfügung. Für die vorliegende Aufgabe ist es nicht wesentlich, dass der korrekte Wörterbucheintrag gefunden wird; vielmehr sollen Wortformen zu demselben Stamm eine Äquivalenzklasse bilden. Ein ernsthafteres Problem ist der Zusammenfall von Formen, die eigentlich getrennt werden müssten (*weiss* 1. Sg. vs. *weiss* Adj.). Es würde aber den Rahmen dieser Arbeit sprengen, eine durchwegs korrekte Lemmatisierung anzustreben, und wir müssen solche Fälle als Störung der Statistik in Kauf nehmen.

⁴zugänglich (Dez. 2003) über ein Html-Interface: <http://www.ifi.unizh.ch/CL/UIS/parser.html>

W	$\text{freq}_{\text{nzz}}(W)$	$\text{freq}_{\text{lit}}(W)$
nathan	2	449
gräfin	2	317
deine	2	257
ward	5	517
wollt	2	157
andre	2	140
bist	6	376
...		
fabel	6	6
glücklichste	2	2
...		
gewöhnlicher	8	8
sowie	1326	6
schweiz	1798	8
april	1643	7
donnerstag	503	2
franken	929	2
zürich	1679	2

Abbildung 3: Listenauszug Frequenzverhältnisse

2.4 Kollokation

Nach Auslassung übermässig häufiger Wörter (*stop words*) betrachte ich jede Kombination von zwei (Nomen + (Verb oder Adjektiv)) aus der Schnittmenge des Vokabular von `lit` und `nzz` in einem Satz prinzipiell als Kandidaten für eine Redensart. Ich komme so auf 487229 unterschiedliche Wortpaare, deren Redensartlichkeit bei jedem ihrer Auftreten im Text beurteilt werden soll.

Zur Veranschaulichung der Verteilung dieser Wortpaare vergleiche als Mass für die Kollokation eines Paares (A, B) die Summe der Kookkurrenzen normiert mit dem geometrischen Mittel der Einzel-Häufigkeiten $(\#A, \#B)$:

$$\text{koll}(A, B) = \sum_{\text{Phrasen}(W_i=A, W_j=B)} \frac{1}{|i - j|^\alpha \sqrt{\#A \#B}};$$

in Abb. 4 für beide Korpora nebeneinander. Der Exponent α in $\text{koll}(A, B)$ soll zeigen, dass die Kollokation mit zunehmender Distanz der Wörter schwächer bewertet wird, aber nicht zwingend mit einer strengen $1/x$ -Gewichtung. Ich setze $\alpha = 0.5$, so dass, wenn drei Wörter eingeschoben werden, die Kollokation noch halb so stark gewichtet wird, als wenn ihre Glieder unmittelbar nebeneinander stehen. Im Zusammenhang mit Redewendungen ist eine solche Gewichtung nach Wortabstand mit Vorsicht zu geniessen, da ein Einschub eines Nebensatzes die Glieder weit auseinanderspreizen kann.

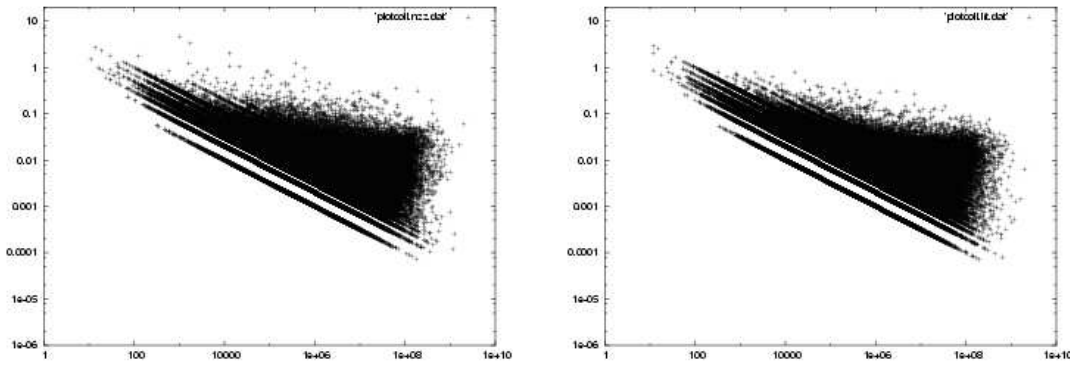


Abbildung 4: Scatterplot kollozierter Wortpaare: waagrecht ist das Produkt der unkorrelierten Häufigkeit, senkrecht die Häufigkeit kollozierten Auftretens (in derselben Phrase, gewichtet nach Anzahl dazwischenstehender Wörter, vgl. Text)

2.5 semantische Felder

Der rechenaufwendigste Teil dieses Ansatzes ist die Bestimmung eines semantischen Feldes für ein bestimmtes Kandidaten-Wortpaar (A, B) . Durch Zeit und Rechenleistung sind hier sehr enge Grenzen gesetzt. Optimal wäre eine Berücksichtigung nicht nur der Wörter, die oft in der Umgebung eines Wortes vorkommen, sondern iterativ auch von Wörtern, die diese wiederum bevorzugt in ihrer Umgebung haben.

Die Beschreibung von semantischen Feldern als Vektoren (*conceptual vectors*) ist ein beliebtes Werkzeug des NLP (z. B. Schütze 1998 [7], Widdows & Peters 2003 [9], Lafourcade 2002 [5], Pado & Lapata 2003 [6]). “Vektoren” bestehen für ein Wort aus Zuordnung eines Winkels zu jedem anderen Wort, der die semantische Distanz beschreibt.

Ich beschränke mich auf die Definition eines semantischen Feldes $S_A(B)$ eines Wortes A als

$$\begin{aligned}
 S_A(B) = & \sum_{\text{Artikel } a \in \text{nzz} : A \in a, B \in a, B \neq A} (\alpha_1) \\
 + & \sum_{\text{Phrasen } p \in \text{lit} : A \in p, B \in p, B \neq A} (\alpha_2) \\
 + & \sum_{\text{Phrasen } p_i, p_j \in \text{lit}; A \in p_i, B \in p_j, 0 < |i - j| < k} (\alpha_3)
 \end{aligned}$$

Zu deutsch: Das Mass der semantischen Zugehörigkeit eines Wortes B zu einem Wort A setzt sich aus drei Komponenten zusammen:

- Kookkurrenz von A und B in einem Artikel (viz. Zeitungsartikel) in **nzz**, gewichtet mit einem Koeffizienten α_1 .
- Kookkurrenz von A und B in einer Phrase in **lit**, gewichtet mit einem Koeffizienten α_2 .
- Okkurrenz von B in einer Phrase in **lit** innerhalb einer Kohärenzlänge von k Phrasen neben einer Phrase, die A enthält, gewichtet mit einem Koeffizienten α_3 .

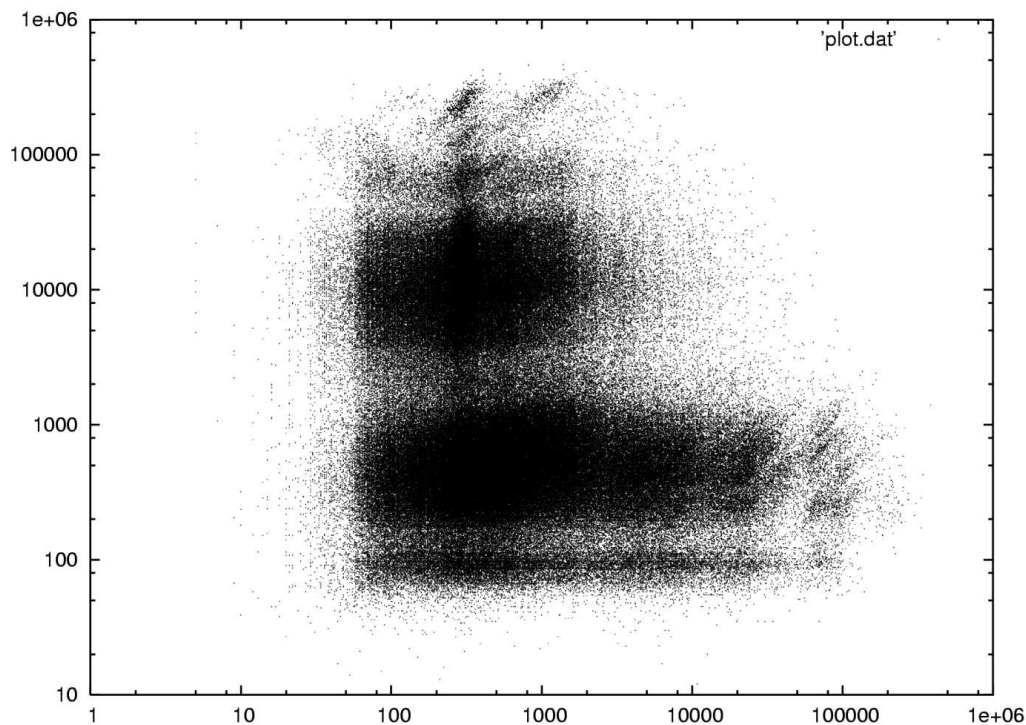


Abbildung 5: Scatterplot von $(S(A_i), S(B_i))$

Ich setze $\alpha_1=1$, $\alpha_2=10$, $\alpha_3=2$, $k = 3$. Der Wert von α_1 ist tief gewählt, einerseits da die ‘klassische’ Semantik von `lit` von Interesse ist und andererseits weil NZZ-Artikel relativ lang sein können, so dass der semantische Zusammenhang nicht mehr gegeben ist; Werte aus dieser Komponente werden dann dominant, wenn eine bestimmte Kombination in `lit` fehlt. Werte $S_A(A)$ können nur durch die dritte Komponente zustande kommen, d. h. durch gehäuftes Auftreten von A in `lit`. Ich verzichte auch auf eine stärkere Gewichtung der Kookkurrenzen in `nzz` wegen der journalistischen Mode des *Anfeuerns*: kurze “szenische Einleitungen” mit hohem Anteil an Kernvokabular, die mit dem Rest des Artikels nicht in direktem semantischem Zusammenhang stehen.

Aus dieser Quantifikation des semantischen Zusammenhangs S lässt sich nun für jedes Wort im Text die Stärke seiner semantischen Bindung an seine Umgebung beziffern. Diese Operation führe ich für jedes der oben bestimmten Kandidaten-Wortpaare durch.

Für ein Wort $W_i \in a$, a ein Artikel in `nzz`, lässt sich nun eine Bindung $S(W_i)$ an den Kontext des Artikels definieren:

$$S(W_i) = \sum_{W_j \in a} \frac{S_{W_j}(W_i)}{\sqrt{\text{freq}_{\text{nzz}} W_j \cdot \text{freq}_{\text{lit}} W_j}}$$

Also die Summe über die semantische Bindung an die einzelnen Wörter des Artikels, gewichtet mit ihrer absoluten Häufigkeit (geometrisches Mittel über die Korpora). Für jede Okkurrenz (A_i, B_i) eines Kandidaten (A, B) lässt sich so also ein Kontext-Wert $(S(A_i), S(B_i))$ berechnen⁵.

⁵Für meine Korpora ein Rechenaufwand in der Größenordnung 100 GHz×h (Programmierung in perl)

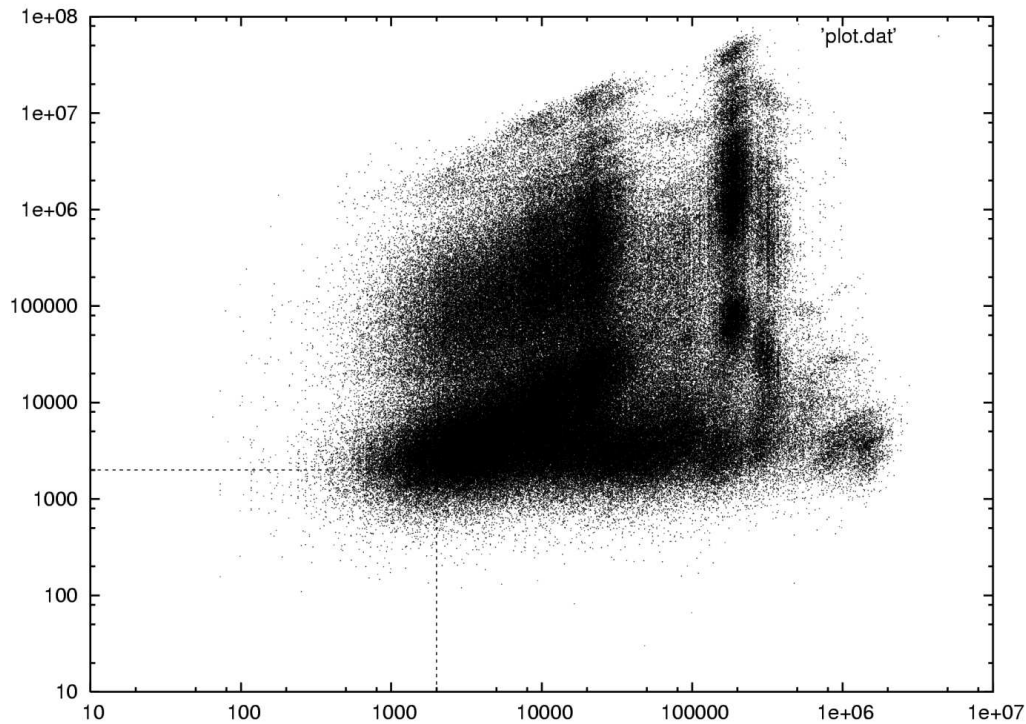


Abbildung 6: $(P(A_i), P(B_i))$ für $n = 80, \alpha = 1$

Abb. 5 zeigt den resultierenden Scatterplot.

Diese Grössen sind noch behaftet mit den absoluten Frequenzen von A und B . Wir wollen diese Werte aber nicht grundsätzlich lossein, sondern wollen verwerfen, dass wir für Redensarten ‘normale’ Wörter erwarten. Ausserdem soll einfließen, dass wir altertümliche Wörter bevorzugen, also solche mit $\text{freq}_{\text{nzz}}W < \text{freq}_{\text{lit}}W$. Betrachte daher

$$P(W) = S(W) \cdot \left(\frac{\text{freq}_{\text{nzz}}W + \text{freq}_{\text{lit}}W}{n} + \frac{n}{\text{freq}_{\text{nzz}}W + \text{freq}_{\text{lit}}W} \right)^\alpha \cdot \frac{\text{freq}_{\text{nzz}}}{\text{freq}_{\text{lit}}W}$$

mit n einer ‘typischen’ absoluten Frequenz und α einem Koeffizienten, der beschreibt, wie stark Wörter mit Frequenzen in der Nähe von n bevorzugt werden sollen. Abb. 6 zeigt $(P(A_i), P(B_i))$ für $n = 80, \alpha = 1$; Links unten stehen die Paare mit schlechter Korrelation mit ihrer Umgebung, rechts oben die Paare mit guter Korrelation. Wir vermuten die Redensarten in der linken unteren Ecke des Plots.

Das asymmetrische Bild kommt dadurch zustande, dass die Paare in der Ordnung $(N, [VA])$ erscheinen, waagrecht ist also der Wert für den nominalen Teil des Paares aufgetragen. Insgesamt erscheinen 393946 Belegstellen als Punkte im Scatterplot.

2.5.1 Schuhe, Teufel

Um ein Gefühl für die errechneten Grössen zu bekommen, greife ich einige Paare (Schuh, B), (Teufel, B) heraus (natürlich im Hinblick auf Redensarten $\{\text{Schuh}, \text{drücken}\}, \{\text{Teufel}, \text{fürchten} [\text{das Weihwasser}]\}$ etc.).

(a)	$(P(\textit{Teufel}), P(\textit{stecken}))$	= (711, 1016)
(b)	$(P(\textit{Teufel}), P(\textit{stecken}))$	= (624, 1449)
(c)	$(P(\textit{Teufel}), P(\textit{erretten}))$	= (1412, 38939)
(d)	$(P(\textit{Teufel}), P(\textit{schimmern}))$	= (552, 1500)
(e)	$(P(\textit{Teufel}), P(\textit{fürchten}))$	= (552, 717)
(f)	$(P(\textit{Schuh}), P(\textit{drücken}))$	= (1230, 694)
(g)	$(P(\textit{Schuh}), P(\textit{drücken}))$	= (1099, 722)
(h)	$(P(\textit{Schuh}), P(\textit{schieben}))$	= (1355, 1905)
(i)	$(P(\textit{Schuh}), P(\textit{schieben}))$	= (1261, 925)
(j)	$(P(\textit{Schuh}), P(\textit{funkeln}))$	= (3776, 1531)

(a) Steckt der Teufel im Detail, oder soll der Improvisationskunst des Einzelnen grösstmöglicher Raum gewährt werden? — (b) Zudem hatte sich bei früheren Aufteilungsversuchen gezeigt, dass der Teufel im Detail steckt und hüben wie drüben wenig Kompromissbereitschaft besteht. — (c) Denn es müsse “uns Christen kein Scherz sondern grosser Ernst sein, hie wieder Rat zu suchen und unser Seelen von den Jüden, das ist: vom Teufel und ewigen Tod, zu erretten”. — (d), (e) Bei beiden Stellungnahmen schimmerte die Erkenntnis durch, dass die IRA bis anhin derartige Unterbrechungen ihrer Kampagne fürchtete wie der Teufel das Weihwasser. — (f) Doch auch da drückt der Schuh. Der Bund kann ja heute nichts anderes mehr tun, als den verarmten Marquis aus Goldonis “Locandiera” nachzuahmen und “seinen Schutz anzubieten”! — (g) Streiks in verschiedenen Braunkohlebergwerken, die am Donnerstag wieder aufgeflammt sind, zeigen indessen, dass die meisten Menschen hierzulande (unabhängig vom Verfassungsmodell) der Schuh zurzeit ganz anderswo drückt. — (h) Der Zuger Polizeidirektor Hanspeter Uster, dessen Partei, der Sozialistisch-Grünen Alternativen, die Störaktion in die Schuhe geschoben wurde, ging demonstrativ selber zu diesem Anlass. — (i) Trotz einem neuen Waffenstillstand, der am Dienstagabend zwischen Vertretern der bosnischen Serben und der Uno-Schutztruppen (Unprofor) in Pale bei Sarajewo vereinbart wurde, sind die Kämpfe in der mit 60 000 bis 70 000 Menschen überfüllten Stadt Gorazde auch am Mittwoch weitergegangen, wobei sich Muslime und Serben gegenseitig die Schuld dafür in die Schuhe schoben. — (j) Glitzernde Kleider liegen auf dem Boden, Schuhe funkeln im Rampenlicht, flippige Hüte scheinen achtlos hingeworfen, und vom Rand blenden die Spiegel, beleuchtet von Scheinwerfern.

Man sieht, wie $P(\textit{Teufel})$ bzw. $P(\textit{Schuh})$ bei wörtlicher Verwendung in (c), (j) deutlich höhere Werte zeigt. Interessant auch die Kopplung von *Teufel* an *schimmern* in (d); beide Wörter waren redensartlich verwendet, aber nicht Teil derselben Redensart; das korrekte (e) $\{(Teufel), P(fürchten)\}$ wird dann auch bevorzugt.

Aufgrund dieser Resultate (an Redensarten beteiligte Wörter erscheinen mit Werten P um 1000) wollen wir versuchsweise Wortpaare mit $P(A) < 2000$, $P(B) < 2000$ als Redensartskandidaten verwenden.

3 Resultate

Betrachten wir den oben angesetzten Schnitt $P(A) < 2000$, $P(B) < 2000$ (eingezeichnet in Abb. 6: In diesen Bereich fallen nur 1.5% der getesteten Wortpaare (die ja auch bereits eine Vorauswahl darstellen, da sie alle zum Vokabular von lit gehören). Wenn sich der überwiegende Teil der Redensarten in diesem Bereich wiederfindet, wäre uns also sicher gelungen, auf die beschriebene Weise aus einer Anfangs-Population einen überwiegenden Anteil als nicht-redensartlich zu erkennen. Ausserdem kann bei Betrachtung eines einzelnen Wortes W (vgl. oben *Teufel*) der Wert von $P(W)$ signifikante Auskunft geben über Bindung an den Text und damit über wörtli-

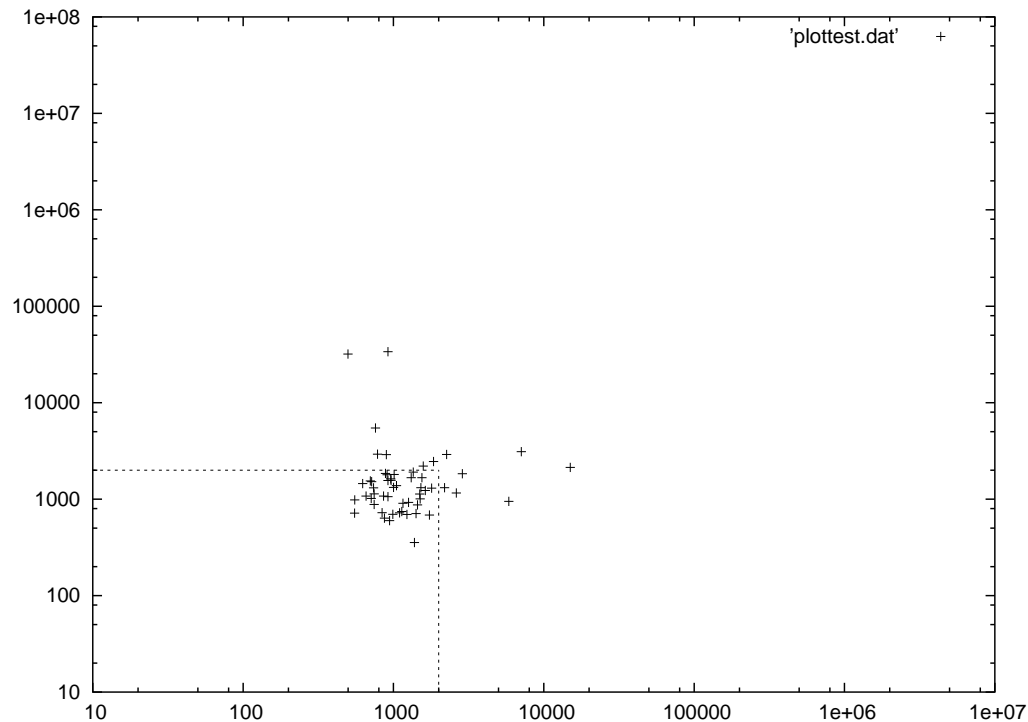


Abbildung 7: $(P(A_i), P(B_i))$ für ausgesuchte Redensarten

che oder übertragene Verwendung. Für eine weitere Beurteilung der verbleibenden Kandidaten wird mit sie gegen einen sehr viel grösseren Korpus auf Reokkurrenz testen.

Eine vollständige Sichtung der Resultate von Hand, mit einer Entscheidung für jedes Wortpaar, wie gut seine Redensartlichkeit erkannt wurde, ist natürlich nicht möglich. Stattdessen evaluieren wir die Aussagekraft des eingeführten *cut* mit der folgenden Auswahl von Redensarten:

‘Tod und Teufel fürchten’, ‘der Teufel soll ...holen’, ‘der Teufel steckt im Detail’, ‘den Finger (auf den wunden Punkt) legen’, ‘einen Finger rühren’, ‘auf die Finger schauen’, ‘der Schuh drückt’, ‘in die Schuhe schieben’, ‘die Achseln zucken’, ‘der Atem stockt’, ‘unter Dach und Fach bringen’, ‘sich ... auf die Fahne schreiben’, ‘das Handwerk legen’, ‘die Katze im Sack kaufen’, ‘dicke Luft’, ‘um die Nase wehen’, ‘auf taube Ohren stossen’, ‘vom Tisch wischen’, ‘Wellen werfen/schlagen’, ‘in die Wüste schicken’.

Eine Bestätigung der Methode anhand dieser Auswahl ist aber nur bedingt aussagekräftig: Wohl habe ich diese Redensarten ‘willkürlich ausgewählt’, aber das unweigerlich vor dem Hintergrund der vorher über Redensarten getroffenen Annahmen.

(Teufel, fürchten) (1444, 873); *(Teufel, fürchten)* (552, 717); *(Teufel, fürchten)* (5849, 948); *(Teufel, holen)* (1312, 1669); *(Teufel, stecken)* (711, 1016); *(Teufel, stecken)* (624, 1449); *(Finger, legen)* (744, 1135); *(Finger, legen)* (701, 1544); *(Finger, legen)* (907, 1792); *(Finger, rühren)* (840, 725); *(Finger, schauen)* (871, 635); *(Schuh, drücken)* (1230, 694); *(Schuh, drücken)* (1099, 722); *(Schuh, schieben)* (1355, 1905); *(Schuh, schieben)* (1261, 925); *(Schuh, schieben)* (2189, 1314);⁶

⁶Die Sprecher der MLNF und die Vertreter der Sicherheitskräfte schieben die Gewaltakte gewöhnlich “Ban-

(*Schuh, schieben*) (2618, 1159);⁷ (*Achsel, zucken*) (1001, 1322); (*Atem, stocken*) (1414, 708); (*Dach, bringen*) (919, 1563); (*Fach, bringen*) (7085, 3112); (*Fahne, schreiben*) (885, 1851); (*Fahne, schreiben*) (960, 1626); (*Fahne, schreiben*) (716, 1516); (*Handwerk, legen*) (737, 1312); (*Handwerk, legen*) (1846, 2453);⁸ (*Katze, kaufen*) (1523, 1315); (*Luft, dick*) (2873, 1837); (*Luft, dick*) (14997, 2131);⁹ (*Luft, dick*) (1378, 355); (*Luft, dick*) (1794, 1300); (*Nase, wehen*) (1137, 744); (*Nase, wehen*) (990, 697); (*Nase, wehen*) (1627, 1232); (*Ohr, taub*) (658, 1075); (*Ohr, taub*) (783, 2931);¹⁰ (*Ohr, taub*) (943, 599); (*Ohr, taub*) (963, 1566); (*Streit, brechen*) (1587, 741); (*Tisch, wischen*) (1047, 1380); (*Tisch, wischen*) (897, 2897);¹¹ (*Tisch, wischen*) (2259, 2903);¹² (*Tisch, wischen*) (1577, 2204);¹³ (*Welle, werfen*) (744, 881); (*Welle, werfen*) (920, 1063); (*Welle, werfen*) (860, 1075); (*Welle, werfen*) (1495, 1132); (*Welle, werfen*) (1507, 1008); (*Welle, schlagen*) (759, 5466);¹⁴ (*Welle, schlagen*) (1546, 1668); (*Welle, schlagen*) (553, 981); (*Welle, schlagen*) (1010, 1805); (*Wüste, schicken*) (1733, 685); (*Wüste, schicken*) (1158, 906); Aus dieser Auflistung – sowie dem zugehörigen Scatterplot (Abb. 7) – ist ersichtlich dass sich die ausgesuchten Redensarten weitgehend ‘erkannt’ wurden. Deutliche Ausreisser sind feststellbar in Sätzen, wo (*Schuhe, schieben*) bzw. (*Luft, dick*) tatsächlich einmal in wörtlichem Kontext auftreten.

3.1 Zusammenfassung

In einem Testkorpus (NZZ) sollte nach statistischen Kriterien nach Redensarten gesucht werden. Dazu wurde ein Kontroll-Korpus (klassische Literatur) herangezogen. Durch Vergleich der Worthäufigkeiten in beiden Korpora wurden Wortpaare als Kandidaten für Redensarten gewonnen. Und mithilfe eines aus beiden Korpora gewonnenen semantischen Feld für jedes Wort wurde für jedes Auftreten eines dieser Wortpaare im Testkorpus eine semantische Bindung an den Text berechnet. Die Erwartung dabei war, dass eine figurative Redensart nur eine schwache semantische Bindung an ihre jeweilige Textumgebung haben wird.

Es ist (erwartungsgemäss) nicht gelungen, Redensarten nach rein statistischen Kriterien

reiten” oder “abtrünnigen Muslimrebelln” wie der “Abu-Sayaff-Gruppe” in die Schuhe, doch es ist durchaus denkbar, dass da und dort die Sicherheitskräfte oder die MLNF ihre Hände im Spiel haben.

⁷Angelockt durch fröhliche Stimmen schiebt man eine Türe beiseite, entledigt sich auf den Steinfliesen des Eingangs seiner Schuhe und betritt eine gemütliche Gaststätte.

⁸Der alte Hunnicutt ist durch den Verkauf von schwarzgebranntem Whisky reich geworden. Als er aber eine Ladung mit giftigem Schnaps ausliefern will, versucht ihm der Schwarzbrenner Harley das Handwerk zu legen.

⁹Das Mittelmeertief, das auf seinem Weg nach Nordosten – es hatte am Samstag Finnland erreicht – vor allem in Ostdeutschland verheerende Niederschläge ausgelöst hatte, hinterliess in den unteren Schichten sehr feuchte Luft, so dass sich am Freitag eine dicke Hochnebeldecke bildete, mit Obergrenze um etwa 1800 Meter.

¹⁰Eine Kritik, die nicht auf taube Ohren stiess, denn nach diversen Investitionen im Zuschauerbereich verbesserte der Reitverein vom Kempttal zum 60-Jahr-Jubiläum der sogenannten Osterrennen die Piste.

¹¹Dies gelang ihr tatsächlich, gleichzeitig zeigten aber die Wahlresultate, dass sich das ethnische Problem damit nicht vom Tisch wischen liess: die Regierungspartei schnitt gegenüber kleineren tamilischen und muslimischen Formationen schlecht ab.

¹²Mit wenigen Worten schienen drei Jahrzehnte der Feindseligkeit vom Tisch gewischt.

¹³Während beispielsweise im Bahnhof Zürich die ursprüngliche Idee einer Glasgestaltung zwischen altem und neuem Teil von der zuständigen Kommission schon bald vom Tisch gewischt wurde und ausgestopfte Vögel das Rennen machten, wird der neue Londoner Flughafen Stansted von einem riesigen Glasfenster von Brian Clarke geprägt.

¹⁴Ein Vorschlag der australischen *Federal Airports Corporation* (FAC), Fluggesellschaft ten, welche die neue, dritte Landepiste in Sydney benützen, mit einer “*Lärmsondersteuer*” zu belasten, hat grosse Wellen geschlagen.

völlig zu isolieren. Die beschriebene Methode hat es hingegen erlaubt, die Kandidaten anhand ihres semantischen Bezugs zu ihrer unmittelbaren Umgebung deutlich einzuschränken (wenige Prozent der ursprünglichen Kandidaten). Für einen gegebenen Ausdruck lässt sich so tatsächlich eine brauchbare Entscheidungshilfe gewinnen, ob er in einem bestimmten Kontext redensartlich verwendet wird¹⁵

Eine weiterführende Arbeit könnte sich mit der Kombination der behandelten Methode mit lexikonbasierten Ansätzen befassen, mit dem Ziel, eine praktisch einsetzbare Bewertung zu erreichen, mit welcher Wahrscheinlichkeit ein gegebener Ausdruck in einem gegebenen Text figurativ oder im Wortsinne verwendet ist.

Literatur

- [1] I. Fischer and M. Keil, *Parsing decomposable idioms*, 1996.
- [2] Ingrid Fischer and Martina Keil, *Von großen Böcken und einer Menge Staub - Zur maschinellen Verarbeitung modifizierter Idiome mit semantisch-autonomen Komponenten*, (1996).
- [3] Catherine L. Harris, *Psycholinguistic Studies of Entrenchment*, 1997.
- [4] Martina Keil, *Wort für Wort. Repräsentation und Verarbeitung verbaler Phraseologismen*, Sprache und Information **35** (1997).
- [5] Mathieu Lafourcade, *Lens effects in autonomous terminology learning with conceptual vectors*, (2002).
- [6] Sebastian Pado and Miralla Lapata, *Constructing Semantic Space Models from Parsed Corpora*, (2003).
- [7] Hinrich Schütze, *Automatic Word Sense Discrimination*, Computational Linguistics **24(1)** (1998), 97–123.
- [8] Martin Volk, *The Automatic Translation of Idioms. Machine Translation vs. Translation Memory Systems*, (1998).
- [9] Dominic Widdows and Stanley Peters, *Quantum Logic of Word Meanings: Concept Lattices in Vector Space Models*, (2003).

¹⁵Es soll noch einmal betont werden, dass in dem hier durchgeführten Versuch alle Wort-Kombinationen über dieselbe Schnur geschlagen werden; wäre die Aufgabe, für einen bestimmten Ausdruck (A, B) bei jedem Vorkommnis über seine Redensartlichkeit zu entscheiden, wäre das Resultat noch *viel besser* (wie oben illustriert an den Fällen (*Schuh, schieben*), (*Luft, dick*): ($P(A_i), P(B_i)$)) müsste nicht an einem absoluten Wert gemessen werden, sondern seine Variation für festes (A, B) wäre Indiz für Redensartlichkeit.