Universität zu Köln
Sprachliche Informationsverarbeitung
Hauptseminar: Angewandte linguistische Datenverarbeitung
Prof. Dr. Jürgen Rolshoven
Hausarbeit

# Semantische Spezifität im Word Space Model

Von C. Friedrich

(Vorgelegt am 2. September 2016)

Ladida, ladidu

# Inhaltsverzeichnis

# 1 Einleitung

In dieser Arbeit verfolge ich zwei Ziele: Zum einen suche ich ein Maß für die semantische Spezifität von Begriffen. Dazu entwerfe ich, ausgehend von einigen Hypothesen, verschiedene Maße und teste diese dann mit Hilfe von Korpora und einem Testsatz von Wortpaaren. Zum anderen wende ich die so erprobten Maße auf die Fragestellung an, wie sich Anglizismen im Deutschen hinsichtlich ihrer semantischen Spezifität von ihren englischen Ursprungswörtern unterschieden. In Kapitel 2 und 3 beschäftige ich mich mit dem ersten Ziel, die Anwendung erfolgt in Kapitel 4.

Ausgehend von folgenden drei Thesen stelle ich drei Maße für die semantische Spezifität vor:

1. Semantisch spezifischere Wörter treten seltener auf.
2. Semantisch spezifischere Wörter stehen mit weniger verschiedenen Wörtern im Kontext
3. Semantisch spezifischere Wörter haben einen Kontext, der sich semantisch ähnlicher ist.

These 1 tritt in dieser Arbeit in Form des Document Frequency Score auf (siehe Abschnitt 2.1), These 2 in Form des Non-Zero Dimensions Score (Abschnitt 2.4) und These 3 in Form des Mean Distance to Centroid Score (Abschnit 2.7).

Die verwendeten Maße sind dabei statistische Maße, also Maße, die irgendetwas mit Zählen zu tun haben (Manning und Schütze, 1999). Konkreter: Maße, die sich auf die Anzahl bestimmter Eigenschaften von Wörtern in Textmengen beziehen. Welche Berechnungen man anschließend mit den gezählten Werten anstellt, legt das mathematische Modell der Arbeit fest. Ich werde dazu das Word Space Model verwenden, das in Abschnitt 2.2 vorgestellt wird.

### Semantische Spezifität

Was ist eigentlich semantische Spezifität? Jones (1972, 11) Beschreibt die Spezifität eines Begriffes so:

> Specificity ... is a semantic property of index terms: a term is more or less specific as its meaning is more or less detailed and precise.

Ein Wort ist also eine spezifischere Bedeutung, wenn es konkreter, detaillierter oder präziser ist. Ich denke, dass man so schnell ein intuitives verständnis davon hat, was mit semantischer Spezifität gemeint ist. Anders formuliert könnte man auch sagen, dass ein Wort spezifischer ist, wenn es sich auf weniger Situationen anwenden lässt. Es schließt mehr Sachverhalte von vornherein aus. Bezogen etwa auf die Tierwelt sind spezifischere Bezeichnungen diejenigen mit der geringeren Extension: Mit 'Säugetier' kann man abertausende verschiedenste Tiere bezeichnen, mit 'Europäischer Feldhamster' muss ich schon einige sehr spezielle Tiere raussuchen, die ich korrekt damit bezeichne.

Weeds et al. (2004) proposed a notion of distributional generality, observing that more general words tend to occur in a larger variety of contexts than more specific words. For

example, we would expect to be able to replace any occurrence of cat with animal and so all of the contexts of cat must be plausible contexts for animal. However, not all of the contexts of animal would be plausible for cat, e.g., “the monstrous animal barked at the intruder”.



# 2 Ein Maß für die semantische Spezifität

Was ist die Challenge des Experiments?

Für die Auswahl eines passendes Word Space Models gibt es verschiedene Optionen. Naheliegend ist etwa eine Dokument-Wort-Matrix, in der für jedes Dokument des Korpus angegeben wird, wie häufig jedes analysierte Wort des KOrpus darin auftritt. Eine solche vollständige Matrix ist eine Darstellung des sogenannten Dokumentenraums (*Document Space*)(Manning und Schütze, 1999, S.296). Ebenfalls naheliegend ist eine Wort-Wort-Matrix, die die Relationen der Wörter im Korpus untereinander einzufangen versucht. Sie stellt den sogenannten Wortraum dar (*Word Space*). Ich werde mich ihn dieser Arbeit statt auf den Dokumentenraum auf Methoden aus dem Wortraum stützen. Schütze und Pedersen (1994) argumentieren dafür, dass sowohl quantitativ als auch qualitativ reichhaltigere semantische Informationen auf Basis des Word Space Models gewonnen werden können.

Wie die Relationen zwischen Wörtern aussehen können und wie man mit ihnen eine Näherung der semantischen Spezifität berechnen kann, möchte ich in diesem Kapitel vorstellen. Dazu entwickele ich verschiedene Berechnungsmethoden und kombiniere diese erst einmal ohne starke theoretische Vorannahmen. Dann stelle ich ein Experiment vor, dass die Berechnungsmethoden auf ihre Tauglichkeit zur Berechnung der semantischen Spezifität prüft. Die vielversprechensten Methoden wende ich dann im anschließenden Kapitel auf die Fragestellung der semantischen Spezifität von Anglizismen an.

## 2.1 Score 1: Document Frequency

Bereits Jones (1972) schlug ein statistisches Maß für die semantische Spezifität eines Wortes vor. Es ist die simple Frequenz, mit der ein Wort im Korpus auftaucht, das ein Indiz für die Spezifität darstellen soll. Caraballo und Charniak (1999) konnten die *document frequency* als Eigenschaft von Worten dazu nutzen, für beliebige Wortpaare festzustellen, welches Wort spezifischer oder genereller ist. Überprüft wurde das mit Beispielwortpaaren, die in einer Hyperonym- bzw. Hyponymrelation zueinander stehen: Das Wort *Getränk* ist ein Oberbegriff zum Wort *Cola*. Für diese Art von Relation gilt: Wenn ein Wort ein Oberbegriff eines anderen ist, so ist der Unterbegriff semantisch spezifischer als der Oberbegriff. Die klar unterschiedene Spezifität ist also eine notwendige Bedingung für die Hyperonym- bzw. Hyponymrelation. Das macht solche Wortpaare zu natürlichen Kandidaten, um Maße für semantische Spezifität zu testen.

Insoweit es die Textgrundlage hergibt, also in Dokumenten geordnet ist, verwende ich die *document frequency* als erste Annäherung an ein brauchbares Maß für die semantische Spezifität. Die Berechnung ist simpel und performant. Entscheidend wird die Frage sein, ob

es den komplexeren Modellen gelingt, eine höhere Erfolgsrate zu erzielen. Daher verwende ich die *document frequency* als Benchmark für die anderen Modelle.

Weil das spätere Ziel ist, Worte in unterschiedlichen Korpora miteinander zu vergleichen, muss die *document frequency* noch normiert werden, wodurch das Maß etwas unabhängiger vom verwendeten Korpus wird.

**Definition.** *Sei $N$ die Gesamtzahl aller untersuchen Dokumente und $df_i$ die Anzahl der Dokumente, in denen das Fokuswort $w_i$ auftritt. Dann ist die* **normierte document frequency** $dfn_i$

$$dfn_i = \frac{df_i}{N}. \tag{1}$$

Hat von zwei Wörtern, die wir miteinander vergleichen wollen das erste eine kleinere $dfn$ als das zweite, ist das nach Caraballo und Charniak (1999) eine gute Heuristik, auch eine höhere semantische Spezifität anzunehmen.

## 2.2 Das Word Space Model

Grundlage dieser Arbeit ist das *Word Space Model* (WSM) oder auch Termvektormodell, das unter anderem sehr ausführlich in Sahlgren (2006) beschrieben wird. Das WSM erhält seine Relevanz in der Computerlinguistik hauptsächlich durch eine zentrale Überlegung:

> **The distributional hypothesis:** *words with similar distributional properties have similar meanings.*

Die Formulierung hier stammt aus Sahlgren (2006, S. 21). Die Idee ist naheliegend: Dem abstrakten Konzept der Bedeutungsähnlichkeit wird durch simple räumliche Nähe repräsentiert. Die statistischen Eigenschaften eines Wortes scheinen nach der Hypothese also auf nicht näher bestimmte mit dem semantischen Inhalt eines Wortes zu korrellieren. Diese Korrelation ist jedoch nicht absolut, sondern steht in Relation zu den Eigenschaften eines anderen Wortes. Das WSM stellt also kein Modell für die absolute Bedeutung eines Wortes dar, man kann jedoch Aussagen über die Bedeutungsähnlichkeit verschiedener Worte treffen. [1]

Zu den statistischen Eigenschaften zählen dabei Phänomene wie die Häufigkeit eines Wortes, die Beziehungen zu anderen Worten in der unmittelbaren Umgebung des Wortes, die Beziehung zu anderen Worten im selben Dokument usw. Das WSM stellt dabei diese Eigenschaften durch Zahlenwerte von verschiedenen Features dar. Ein solches Feature wäre beispielsweise die Häufigkeit des Auftretens eines bestimmten Wortes in direkter Nachbarschaft zum Fokuswort. Die Auswahl dieser Featuremenge legt dabei in sehr relevantem Maße die Aussagen fest, die sich mit Hilfe des WSM treffen lassen. Listet man alle Features eines Fokuswortes auf, so erhält man den Featurevektor des Wortes. Dieser Vektor repräsentiert damit die statistischen Eigenschaften des Fokuswortes im Kontext des Modells, das man ausgewählt hat. Die Ähnlichkeit der Featurevektoren lässt dann Rückschlüsse auf die Bedeutungsähnlichkeit der Worte zu, so die Hypothese.

[1] Je nach Bedeutungstheorie ist das mehr oder minder plausibel. Versteht man Bedeutung primär als Referenz (insbesondere extralinguistisch), so kann diese Analogie nicht viel leisten. Ist vielmehr der Gebrauch des Wortes in der Sprache gefragt, entspricht die Analogie je nach Wahl des konkreten Modells zum Teil sehr deutlich dem Begriff der Bedeutung.

Um die Ähnlichkeit numerisch bestimmen zu können, braucht es für einen solchen Vektorraum eine Methode, die Distanz zwischen den einzelnen Featurevektoren zu bestimmen. Welche davon sinnvoll eingesetzt werden können, wird in den nächsten Abschnitten beschrieben.

## 2.3 Satzkookkurrenzen vs. Fensterkookkurrenzen

Zunächst müssen die Features, welche die Vektoren ausmachen, festgelegt werden. Ein naheliegender Kandidat sind hier diejenigen Wörter, die mit dem Fokuswort in einer bestimmten Art und Weise gemeinsam auftreten, also kookkurrieren. Über einen gesamten Korpus legen diese Wörter den Kontext des Fokuswortes fest. Nun gibt es mehrere Möglichkeiten, diesen Kontext festzulegen. In dieser Arbeit habe ich die folgenden beiden Ansätze gewählt:

**Satzkookkurrenzen** sind diejenigen Wörter, mit denen das Fokuswort gemeinsam in einem Satz auftritt. Gezählt werden dabei für zwei Wörter vorranging die Anzahl der Sätze, in denen die Wörter gemeinsam auftreten. Den Satzkontext des Fokuswortes bilden dann die Menge aller Wörter, die mit dem Fokuswort mindestens einmal gemeinsam in einem Satz auftreten.

**Fensterkookkurrenzen** sind diejenigen Wörter, mit denen das Fokuswort innerhalb eines Fensters von festgelegter Größe gemeinsam auftritt. In Abgrenzung zur Satzkookkurrenz habe ich hier kein symmetrisches Fenster untersucht, sondern nur die nachfolgenden Wörter betrachtet, das rechte, gerichtete Kontextfenster. Die Auswahl der Größe des Fensters ist ebenfalls relevant und führt zu signifikanten Unterschieden.

Der Featurevektor des Fokusworts besteht also aus der Kookkurrenz mit allen anderen Worten des Korpus.

**Beispiel.** Der Beispieltext

> The optional plotz says to frobnicate the bizbaz first. Return a foobang.

soll analysiert werden mit *plotz* als Fokuswort[2]:

| | optional | plotz | says | frobnicate | bizbaz | first | Return | foobang |
|---|---|---|---|---|---|---|---|---|
| plotz | $k_1$ | $k_2$ | $k_3$ | $k_4$ | $k_5$ | $k_6$ | $k_7$ | $k_8$ |

Die Werte $k_1$ bis $k_8$ sind die Ausprägungen der Relation zwischen den jeweiligen Wörtern. Die Ausprägung hängt davon ab, (i) ob Satzkookkurrenz oder Fensterkookkurrenz verwendet wird und (ii) welches Maß zur Berechnung der Kookkurrenz verwendet wird. Die verschiedenen Maße, die ich verwende, werden in Abschnitt 2.5 vorgestellt.

Paradigmatischer vs. Syntagmatischer Kontext

## 2.4 Score 2: Anzahl der Kookkurenzen

Ausgehend vom vorherigen Abschnitt lässt sich eine Wort-Wort-Kookkurrenzmatrix erstellen. Diese Matrix enthält alle Featurevektoren jedes einzelnen Wortes des Korpus als Reihe.

[2] Stopwords wurden bereits entfernt.

Die Matrix ist dabei zwingend quadratisch, aber nicht unbedingt symmetrisch, eben im Falle der Fensterkookkurrenzen.

Eine grundlegende These dieser Arbeit ist, das der Kontext eines Wortes als Indiz für seine semantische Spezifität herangezogen werden kann. Nicht nur die reine Häufigkeit eines Wortes ist entscheidend, sondern auch, mit wie vielen verschiedenen Worten das Fokuswort in Kookkurrenz steht. Beispiel: Ein allgemeines Wort kommt etwas seltener vor als ein Spezielleres, der Kontext des spezielleren Wortes ist jedoch beschränkter als der des Allgemeineren, das mit vielen verschiedenen Worten kookkurriert. In so einem Fall würde die *document frequency* fälschlicherweise das allgemeinere Wort als spezieller auszeichnen.

Die Idee dieses Maßes ist es daher, einfach zu zählen, mit wie viele von Null verschiedene Einträge der Featurevektor des Fokuswortes hat, mit anderen Worten, mit wie vielen verschiedenen Worten das Fokuswort in Kookkurrenz steht.

**Definition.** *Sei $N$ die Gesamtzahl aller (unique) Worte im Korpus und $n_i$ die Anzahl aller (unique) Worte, mit denen das Wort $w_i$ in Kookkurrenz steht, d.h. an dessen Eintrag der Featurevektor von $w_i$ einen von Null verschiedenen Wert aufweist. Dann ist der* **Non-Zero Dimension Score** $nzds_i$

$$nzds_i = \frac{n_i}{N}. \tag{2}$$

Meine These: Ein kleiner $nzds$-Wert eines Wortes spricht für eine höhere semantische Spezifität, ein größerer Wert dafür, dass das Wort eher genereller ist.

## 2.5 Maße für die Kookkurrenzen

Zusätzlich zur Auswahl der Art der Kookkurrenz muss noch ein Maß zur Bestimmung der Kookkurrenz gewählt werden. In dieser Arbeit habe ich dafür vier verschiedene Maße herangezogen. Ich gebe hier die Maße für den Fall der Fensterkookkurenzen an. Für die Satzkookkurrenzen ergeben sich leicht andere Maße, auch weil die resultierende Matrix symmetrisch ist. Bei direktionalen Kontextfenstern gilt das nicht notwendigerweise (und praktisch fast nie). Daher ist bei jedem Maß zu beachten: $\text{score}_{ij}$ ist nicht zwingend gleich $\text{score}_{ji}$.

### 2.5.1 Binäre Kookkurrenz

Die binäre Kookkurrenz zeigt an, ob ein Wort mit einem anderen Wort im gesamten Kontext mindestens mit einer bestimmten Frequenz in Kookkurrenz steht.

Sei $K_{ij}$ die Menge aller Kontextfenster für Wort $w_i$, in denen das Wort $w_j$ auftritt, und $m$ die Mindestanzahl an Kookkurrenzen. Dann ist die **binäre Kookkurrenz**

$$bin_{ij} = \begin{cases} 1 & \text{Falls } |K_{ij}| \geqslant m. \\ 0 & \text{Falls nicht.} \end{cases}$$

Es wäre also $bin_{ij}$ gleich 1 genau dann wenn $w_j$ in mindestens $m$ Kontextfenstern von $w_i$ auftritt.

### 2.5.2 Frequenz der Kookkurrenz

Die Frequenz der Kookkurrenz zählt jedes Auftreten der Kookkurenzen [3].

Sei $K_{ij}$ die Menge aller Kontextfenster für Wort $w_i$, in denen das Wort $w_j$ auftritt, wobei $|X|$ die Kardinalzahl der Menge $X$ bezeichnet. Dann ist die **Frequenz**

$$freq_{ij} = |K_{ij}|.$$

### 2.5.3 Dice-Koeffizient

Der Dice-Koeffizient zieht auch in Betracht, wie häufig die Worte im Korpus generall auftreten. Ein sehr häufiges Wort etwa steht mit vielen anderen Wörtern in Kookkurrenz, ohne dass dieses Auftreten besonders signifikant sein müsste. Der Dice-Koeffizient „bestraft“ solche Wörter, indem durch die Gesamtzahl der Vorkommnisse geteilt wird (Manning und Schütze, 1999, S. 299).

Seien $K_{ij}$ die Menge aller Kontextfenster für Wort $w_i$, in denen das Wort $w_j$ auftritt und $K_i$ und $K_j$ die Mengen aller Kontextfenster, in denen die Wörter $w_i$ resp. $w_j$ auftreten. Dann ist der **Dice-Koeffizient**

$$dice_{ij} = \frac{2|K_i \cap K_j|}{|K_i| + |K_j|}.$$

### 2.5.4 Chi-Square

Pearson's Chi-Square-Test ist etwas elaborierter. Er stellt ebenfalls ein Maß für die Signifikanz der Kookkurrenz zur Verfügung, indem er das tatsächlich beobachtete Auftreten von Kookkurrenzen dem statistisch Erwartbaren gegenüberstellt. Statistisch erwartbar bedeutet dabei die rein zufällige Häufigkeit des gemeinsamen Auftretens in Kookkurrenz, wenn beide Worte unabhängig voneinander wären, d.h. kein Zusammenhang zwischen ihnen bestünde. Der Chi-Square-Wert steigt, je signifikanter die Kookkurrenz.

Zur Berechnung wird zunächst eine Kontigenztabelle erstellt:

| | $w_i$ | $\neg w_i$ |
|---|---|---|
| $w_j$ | Anzahl $w_i$-Fenster mit $w_j$ | Anzahl $w_j$-Fenster ohne $w_i$ |
| $\neg w_j$ | Anzahl $w_i$-Fenster ohne $w_j$ | Anzahl Fenster ohne $w_i$ und $w_j$ |

Sei $T$ die Kontingenztabelle zweier Wörter $w_i$ und $w_j$, $O_{k,l}$ der Wert der Kontingenztafel in Zelle $k, l$ und $E_{k,l}$ der erwartete Wert für Zelle $k, l$, dann ist der **Chi-Square-Wert**

$$chi_{ij} = \sum_{k,l} \frac{(O_{k,l} - E_{k,l})^2}{E_{k,l}}.$$

Für eine detailliertere Beschreibung inklusive Berechnung des erwarteten Werts siehe Manning und Schütze (1999, S. 169ff.).

[3] Eigentlich bezeichnet 'Frequenz' die relative Häufigkeit. Da aber in der Literatur 'Document Frequency' häufig als Anzahl der Dokumente verstanden wird und ich den Term hier auch so benutze, nenne ich der Konsistenz halber diesen Wert auch Frequenz.

## 2.6 Distanzmaße im Vektorraum

Um die Unterschiede der Featurevektoren im Word Space Model zu berechnen, habe ich unterschiedliche Distanzmaße verwendet.

### 2.6.1 Standarisierte Euklidische Distanz

Die normale euklidische Distanz ist ein gutes erstes Maß für die Distanz. Allerdings wird hier jede Dimension des Raumes gleich gewichtet. Gibt es eine sehr große Varianz in dieser einen Dimension, so hat diese Dimension eine unverhältnismäßig große Auswirkung auf die Gesamtdistanz. Das ist z.B. der Fall, wenn physikalische Eigenschaft unterschiedlicher Einheiten in einem Featurevektor zusammengefasst werden. Das ist im Kookkurrenzraum zwar nicht der Fall, trotzdem können so einzelne sehr große Unterschiede unverhältnismäßig viel Ausschlag geben. Um diesen Effekt zu mindern, kann das Maß standarisiert werden. Dazu wird mit der reziproken Varianz des jeweiligen Features multipliziert.

Seien $u$ und $v$ die Featurevektoren der Worte $w_i$ und $w_j$, $u_k$ der $k$-te Wert des Featurevektors $u$. Dann ist die **Standardized Euclidean Distance**

$$se(x_i, x_j) = \sqrt{\sum_k \frac{1}{V_i}(u_k - v_k)^2}.$$

### 2.6.2 Kosinusdistanz

Die Kosinusdistanz liefert ein bereits normalisiertes Maß dafür, wie stark zwei Vektoren korrellieren (Manning und Schütze, 1999, S. 300). Sie verhält sich analog zur euklidischen Distanz für normalisierte Vektoren. Zu beachten ist, dass oben die *Standardized*, nicht die normalisierte euklidische Distanz beschrieben wird.

Seien $u$ und $v$ die Featurevektoren der Worte $w_i$ und $w_j$, wobei $u \cdot v$ das Skalarprodukt von $u$ und $v$ bezeichnet und $|x|$ die Länge eines Vektors $x$. Dann ist die **Kosinusdistanz**

$$cos(x_i, x_j) = \frac{u \cdot v}{|u||v|}.$$

## 2.7 Score 3: Semantische Nähe des Kontextes

Welche Informationen über semantische Sepzifität lassen sich nun aus der so gewonnenen Wort-Wort-Kookkurrenzmatrix gewinnen? Ein naheliegender Ansatz wäre es, die Distanz zwischen den Featurevektoren derjenigen Wörter zu berechnen, die wir vergleichen wollen. Das Resultät wäre eine einzige Zahl, die Auskunft über die semantische Nähe beider Wörter gibt. Daraus lässt sich jedoch nicht auf die semantische Spezifität schließen. Ist die Distanz hinreichend klein, so haben die Wörter anscheinend ähnlichen semantischen Gehalt, ist die Distanz sehr groß, unterscheiden sie sich. Welches der Wörter ist aber nun das semantisch spezifischere?

Was also scheinbar gefordert ist, ist ein Maß pro Wort für die semantische Spezifität, das sich dann mit dem Maß des anderen Wortes vergleichen lässt.

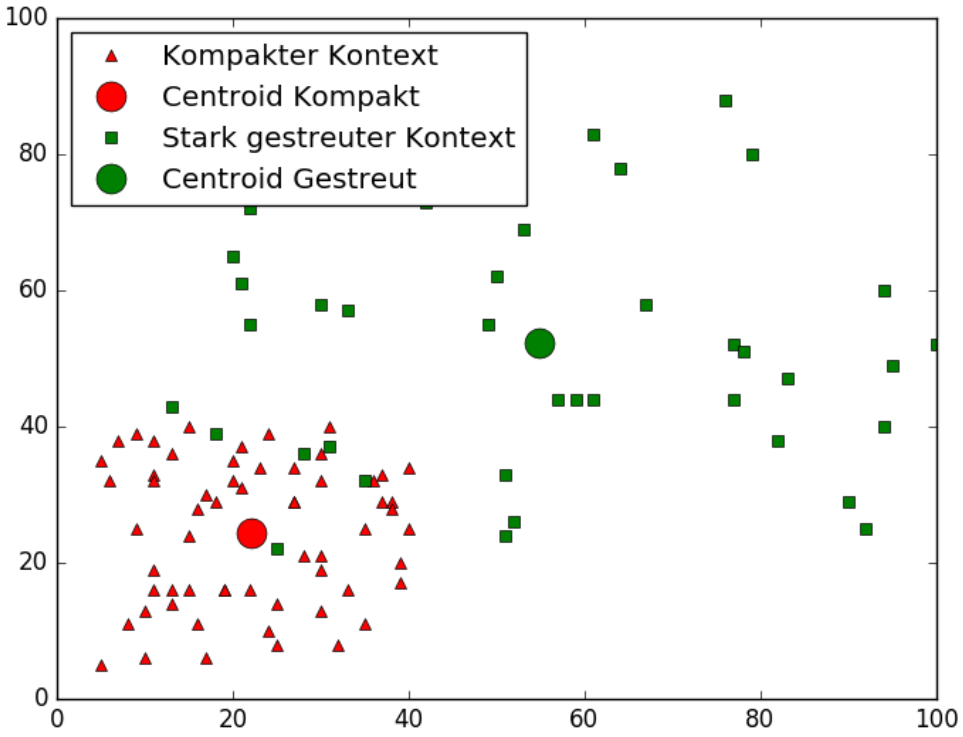


Abbildung 1: Streuung und Kompaktheit verschiedener Kontexte. Die beliebig dimensionalen Featurevektoren werden hier schematisch als Punkte im 2dimensionalen Raum dargestellt. Die Kontexte sind jeweils alle Featurevektoren des Kontextes eines Wortes.

Ist zur Bestimmung der Spezifität eines Wortes nur der direkte Kontext relevant, oder könnte es sein, dass auch die statistischen Eigenschaften der Wörter des Kontextes eine Rolle spielen? Wenn ein Wort mit eher vielen verschiedenen Worten im Kontext steht, die Kontextworte sich untereinander allerdings semantisch sehr ähneln, scheint das ein Wort mit hoher semantischer Spezifität zu sein. Steht ein Wort stattdessen mit eher wenigen Worten im Kontext, diese Worte sind jedoch völlig unterschiedlich aus verschiedensten Gebieten, scheint das ein Wort mit geringer semantischer Spezifität zu sein (vgl. Abbildung 1).

Angenommen, die statistischen Eigenschaften des Kontextes sind relevant für die Spezifität, dann könnte die semantische Nähe der Kontextworte zueinander ein sehr aussagekräftiges Maß für die semantische Spezifität eines Begriffes sein.

Wie ließe sich das in ein Modell übersetzen?

Der Featurevektor des Fokuswortes beschreibt, mit welchen anderen Worten das Fokuswort in Kookkurrenz steht. Für jedes dieser Wörter des Kontextes gibt es nun wiederum einen Featurevektor, der die statistischen Eigenschaften des Wortes beschreibt. Über die distributional hypothesis des Word Space Models lässt sich nun argumentieren, dass geringe intrakontextuelle Distanz im Vektorraum eine hohe semantische Nähe des Kontextes indiziert, und damit einen Hinweis auf die semantische Spezifität des Fokuswortes geben.

Ein geeignetes Maß für die Kompaktheit des Kontextes zu finden, ist also entscheidend.

Hierbei lässt sich ausnutzen, dass die Menge an Featurevektoren, die über den Kontext eines Wortes festgelegt wird, die Form eines *Clusters* annimmt. In der Literatur finden sich einige Ansätze zur Evaluierung von verschiedenen Clusteralgorithmen (Zitat Zitat Zitat). Für dieses Problem sind besonders Maße für die Kompaktheit eines einzelnen Clusters interessant, ohne dabei andere Cluster zu berücksichtigen. Ich möchte also für zwei zu vergleichende Wörter nicht wissen, wie gut ihre Kontexe in Cluster aufgeteilt wurden oder wie sehr die Cluster überlappen (obwohl das sicher auch einige interessante Informationen über die statistischen Eigenschaften der Kontexte liefern würde), sondern evaluieren, wie dicht oder kompakt die jeweiligen Cluster sind. Das Maß soll dann möglichst vom verwendeten Featureraum abstrahieren und die Kompaktheit von Clustern über verschiedenen Vektorräume miteinander vergleichbar machen.

Ein Konzept, das aus der Clusterevaluierung nutzbar gemacht werden kann, ist der *Centroid* bzw. geometrische Schwerpunkt, der den Mittelpunkt aller Featurevektoren repräsentiert. Wie weit sind die einzelnen Featurevektoren vom Schwerpunkt entfernt? Hohe Distanz lässt auf eine weite Streuung schließen, niedrige Distanz auf Kompaktheit. Um das Maß nicht allzusehr zu verkomplizieren, habe ich zur Bestimmung der Verteilung der Distanz der einzelnen Featurevektoren zum Schwerpunkt den Durschnitt aller Distanzen gewählt.

Das vorgeschlagene dritte Maß für die semantische Spezifität eines Begriffes berechnet sich also wie folgt:

**Definition.** *Seien $X_i$ die Menge aller Featurevektoren derjenigen Wörter, mit denen das Wort $w_i$ in einer Kontextrelation steht (s.o.), und $c_i$ der geometrische Schwerpunkt von $X_i$. Dann ist der* **Mean Distance to Centroid Score** $mdcs_i$

$$mdcs_i = \frac{1}{|X_i|} \sum_{x_j \in X_i} dist(x_j, c_i). \tag{3}$$

Die Menge der Wörter des Kontextes werden bei diesem Maß durch jeden von Null verschiedenen Wert im Featurevektor des Fokuswortes bestimmt. Um dem Umstand Rechnung zu tragen, dass die Kookkurrenz des Fokuswortes mit jedem Wort des Kontextes unterschiedlich ausgeprägt ist, verwende ich zusätzlich ein abgewandeltes Maß, das den kompletten Featurevektor des Kontextwortes mit der Ausprägung des zugehörigen Eintrags im Featurevektor des Fokuswortes skaliert.

**Definition.** *Seien $X_i$ die Menge aller Featurevektoren derjenigen Wörter, mit denen das Wort $w_i$ in einer Kontextrelation steht (s.o.), $c_i$ der geometrische Schwerpunkt von $X_i$ und $a_{ij}$ der Eintrag des Featurevektors von Wort $w_i$ von Kontextwort $w_j$. Dann ist der* **Scaled Mean Distance to Centroid Score**

$$sca\_mdcs_i = \frac{1}{|X_i|} \sum_{x_j \in X_i} a_{ij} dist(x_j, c_i).$$

Damit ist $mdcs$ also nur ein Spezialfall von $sca\_mdcs$ mit $a_{ij} = 1$ für alle $i, j$.

Ein Ziel dieses Maßes ist es, das Resultat unabhängiger vom verwendeten Korpus zu machen. Ein Beispiel zur Verdeutlichung: Ein Korpus enthält ein Wort mit einer bestimmten *Document Frequency*. Nun werden dem Korpus eine Menge von Texten hinzugefügt, die

jedoch völlig andere Themengebieten behandeln und das Fokuswort nicht enthalten. Die *Document Frequency* nimmt in starkem Maße ab. Angenommen der Kontext des Wortes ist auch nicht sonderlich stark in den hinzugefügten Texten vertreten, dann steigt die *Mean Distance to Centroid* nicht oder nicht signifikant.

# 3 Semantische Spezifität: Experiment

## 3.1 Textgrundlage

Welche Korpora werden verwendet?
Wie sieht das Präprozessieren aus?
Keine Bigramme.
Frequency Threshold.

## 3.2 Getestete Maße

Im Experiment habe ich die vorgestellten Maße miteinander kombiniert. Maße mit einigermaßen akzeptablen Resultaten möchte ich hier vorstellen. Zur Referenz hier alle relevanten getesteten Maße zur semantischen Spezifität. In der letzten Spalte der Tabelle steht die maximal erreichte Präzision. Für eine genaue Beschreibung siehe Abschnitt 3.3.

**Kookkurrenzmaße**:

- Binär (Abschnit 2.5.1)
- Frequenz (Abschnitt 2.5.2)
- Dice-Koeffizient (Abschnitt 2.5.3
- Chi Square (Abschnitt 2.5.4)

**Maße zur semantischen Spezifität (Scores)**:

- Document Frequency (df, Abschnitt 2.1)
- Non-Zero Dimensions (nzd, Abschnitt 2.4)
- Mean Distance to Centroid (mdc, Abschnitt 2.7)

**Distanzmaße**:

- Standarisierte Euklidische Distanz (Abschnitt 2.6.1)
- Kosinusdistanz (Abschnitt 2.6.2)

## 3.3 Aufbau des Experiments

Um die Validität der dargestellten Maße bewerten zu können, verwende ich eine Menge von Wortpaaren, bei denen offensichtlich ist, welches Wort die höhere semantische Spezifität besitzt. Sehr gute Kandidaten für diese Wortpaare sind Oberbegriffe und Unterbegriffe, also Begriffe, die andere Begriffe klassifizieren oder subsumieren. Diese Relation zwischen Begriffen impliziert, dass der Oberbegriff genereller und der Unterbegriff spezifischer ist. Eine

| Kontext | K.-Maß | Score | Distanzmaß | Slug | Präzision |
|---|---|---|---|---|---|
| - | - | df | - | df | 0.79 |
| Satz | - | nzd | - | sent_nzds | 0.77 |
| Satz | Dice | mdc | Kosinus | sent_dice_mdcs_cosi | 0.80 |
| Satz | Dice | mdc | Euklidisch | sent_dice_mdcs_eucl | 0.80 |
| Satz | Frequenz | mdc | Kosinus | sent_dice_mdcs_cosi | 0.80 |
| Satz | Frequenz | mdc | Euklidisch | sent_dice_mdcs_eucl | 0.80 |
| Fenster | - | nzd | - | win_nzds | 0.80 |
| Fenster | Binär | mdc | Kosinus | win_bin_mdcs_cosi | **0.81** |
| Fenster | Binär | mdc | S-Euklidisch | win_bin_mdcs_seuc | 0.79 |
| Fenster | Frequenz | mdc | Kosinus | win_freq_mdcs_cosi | **0.83** |
| Fenster | Frequenz | mdc | S-Euklidisch | win_freq_mdcs_seuc | 0.80 |
| Fenster | Frequenz | scaled mdc | S-Euklidisch | win_freq_sca_mdcs_seuc | 0.79 |
| Fenster | Dice | mdc | Kosinus | win_dice_mdcs | **0.81** |
| Fenster | Chi Square | mdc | Kosinus | win_chi_mdcs | **0.82** |

vollständige Liste der verwendeten Wortpaare ist Caraballo und Charniak (1999) entnommen und findet sich im Anhang.

Die verschiedenen Maße lasse ich nun zu jedem Wort der Wortpaare bestimmen und vergleiche anschließend, welchen Begriff das Maß als semantisch spezifischer einstuft. Entspricht die Einstufung der intuitiven Vorannahme, das der Unterbegriff semantisch spezifischer ist, wird das als richtige Einstufung gewertet. Die Präzision für ein Maß ist dann einfach der Anteil an geprüften Wortpaaren, den das Maß gemäß der intuitiven Vorannahme richtig einstuft.

Von besonderem Interesse im Falle der Kontextfenster ist dabei die Größe des Kontextfensters. Sahlgren (2006, S. 68) findet in seinen Experimenten ein optimales Kontextfenster von der Größe 2+2, also zwei Wörter zu jeder Seite des Kontextwortes. Jedoch verweist er auch auf Miller und Leacock (2000), die betonen, dass unser Verständnis von adequaten Kontexten noch nicht perfekt ist und man nicht von vornherein ausschließen sollte, dass sich Kontexte anders verstehen lassen, um wertvolle semantische Informationen zu gewinnen. Ausgehend von dieser Idee habe ich die Resultate als Funktion der Größe des Kontextfensters dargestellt (s. Abbildungen 3, 4 und 5.) und bin ohne theoretische Vorannahmen an diese Fragestellung herangegangen.

### 3.4 Resultate

Für alle Kontextfenstergrößen ist die Document Frequency offensichtlich konstant. Zum Vergleich habe ich sie als gestrichelte Linie in jedes Diagramm eingezeichnet.

Der Non Zero Dimension Score ist weiterhin für jedes Kookkurrenzmaß identisch. Zum besseren Vergleich habe ich diesen Score auch in jedes Diagram eingezeichnet (rot).

Der ndzs performt dabei weitesgehend vergleichbar mit dem dfs, mit kleineren Schwankungen abhängig von der Fenstergröße.

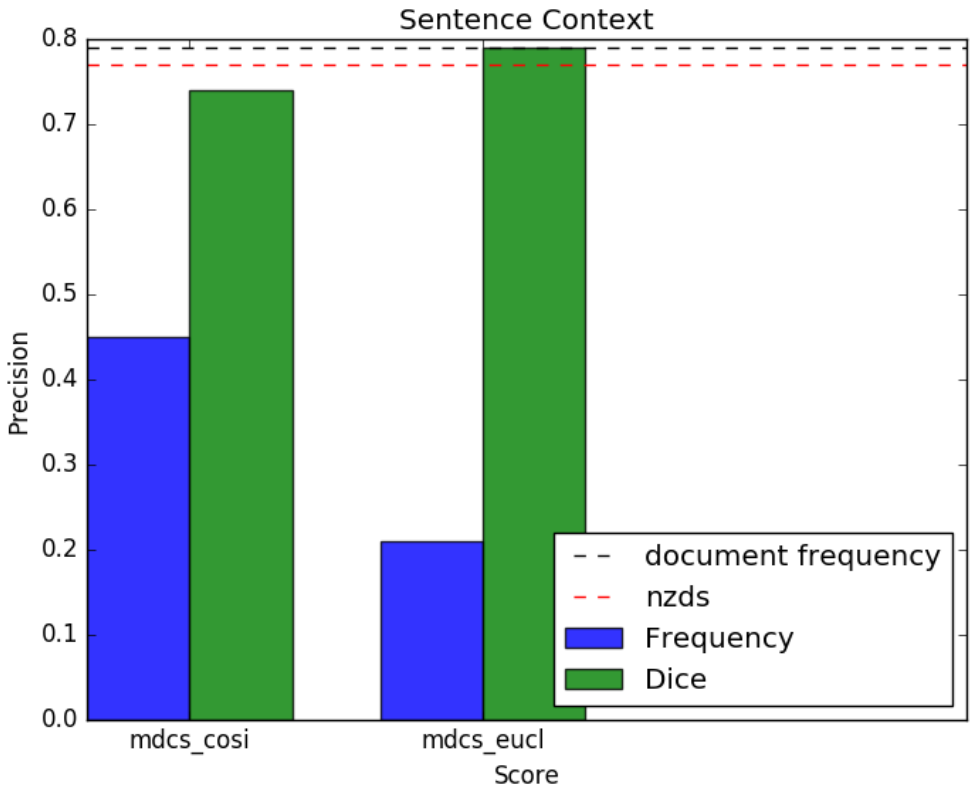


Abbildung 2: Präzision der Maße mit Frequenz und Dice-Koeffizient im Satzkontext

### 3.4.1 Satzkontext

Die Ergebnisse der Analyse nach Satzkontext ist ernüchternd (siehe Abbildung 2). Weder nzds noch mcds mit Kosinusmaß und standarisierter euklidischer Distanz erreichen die Präzisionswerte des Vergleichswert document frequency. Lediglich der Rückschritt auf einfache euklidische Distanz kann an die Vergleichswerte heranreichen. Dann jedoch gibt man die Vergleichbarkeit über verschiedene Vektorräume auf.

### 3.4.2 Fensterkontext, Binäres

Bemerkenswert ist hier, dass sich Mdcs mit Kosinusmaß und Mdcs mit euklidischer Distanz völlig unterschiedlich verhalten (siehe Abbildung 3): Die Präzision bei Kosinusdistanz nimmt mit steigendem Kontextfenstergröße zu, die Präzision der euklidischen Distanz nimmt rapide ab. Interessanterweise ist die Präzision nur bei Kontextgröße $>80$ wirklich besser als der Vergleichswert document frequency.

### 3.4.3 Fensterkontext, Frequency

Es ergibt sich ein ähnliches Bild zum binären Maß (siehe Abbildung 4). Auffällig ist jedoch, dass etwas früher, bei Fenstergröße 60, der mdcs mit Kosinusdistanz ein klar besseres Ergebnis als der Vergleichswert liefert. Der skalierte mdcs erreicht bei keiner getesteten Fesntergröße den Vergleichswert.

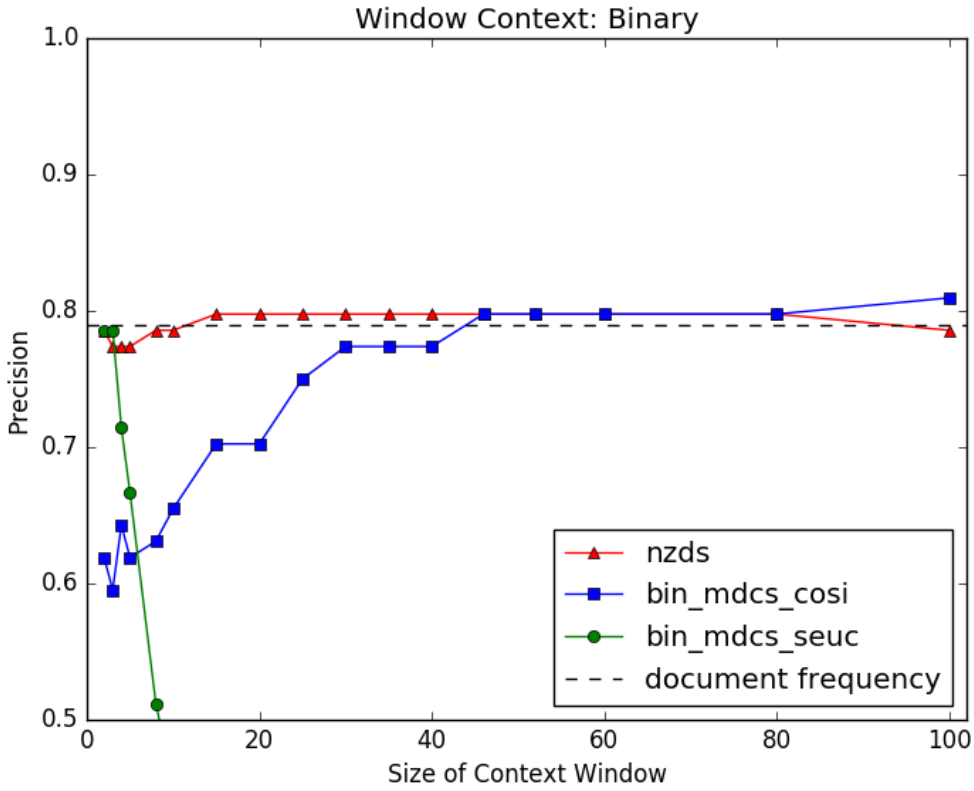


Abbildung 3: Präzision der Maße mit binärem Kookkurrenzmaß im Fensterkontext über Größe des Fensterkontextes

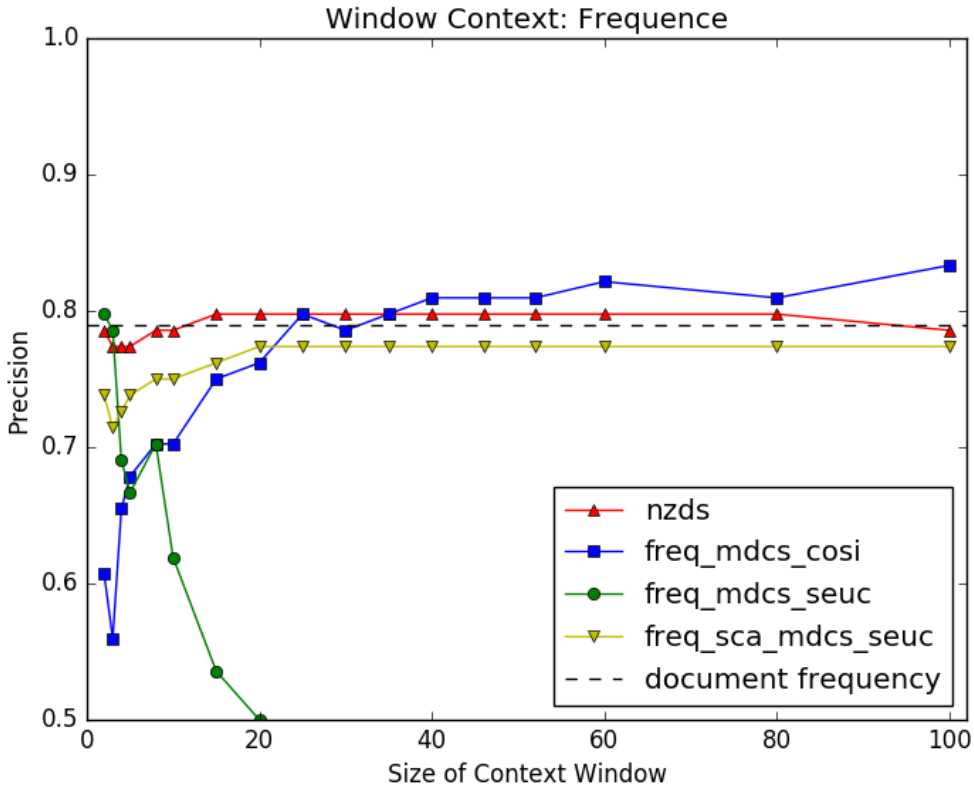


Abbildung 4: Präzision der Maße mit Frequency Kookkurrenzmaß im Fensterkontext über Größe des Fensterkontextes

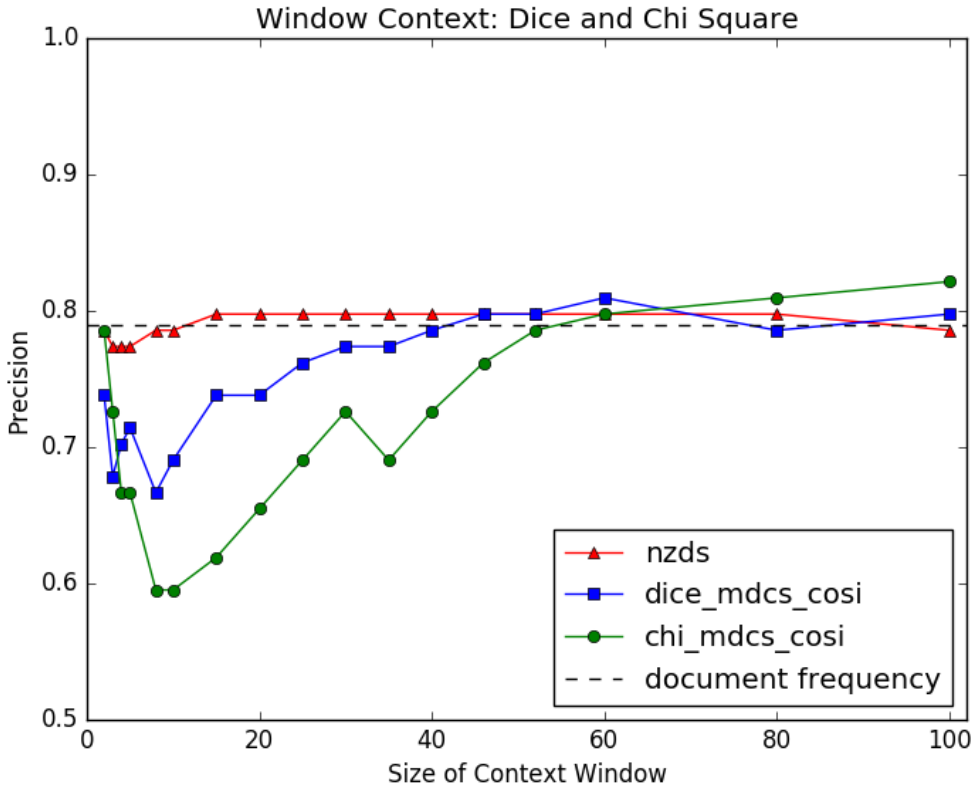


Abbildung 5: Präzision der Maße mit Dice bzw. Chi Square Kookkurrenzmaß im Fensterkontext über Größe des Fensterkontextes

### 3.4.4 Fensterkontext, Dice Koeffizient und Chi Square

Auch hier zeigt sich der für sehr kleine Fenstergrößen akzeptable Präzisionswert und das ausgeprägte Tief im Bereich kleiner 30 (siehe Abbildung 5). Auch auffällig, dass erst bei sehr großem Fensterkontext die Maße eine geeignete Präzision erreichen.

## 3.5 Evaluation

Sehr überraschend ist bei diesen Ergebnissen, dass sich mit einigen Scores bessere Resultate mit sehr großen Kontextfenstern erreichen lassen. Zunächst bestand die Befürchtung, dass kein Maß an die Performanz der simplen Heuristik der document frequency herankommt. Die Ergebnisse zeigen allerdings, dass der hier vorgestellt Ansatz zumindest Potential hat, semantische Spezifität numerisch besser zu fassen als die *document frequency*. In der sehr starken Simplifizierung in den Vorannahmen sind noch einige Verbesserungsmöglichkeiten enthalten, ebenso in der Auswertung der Ergebnisse, mehr dazu in Kapitel 6.

# 4 Anwendung: Anglizismen

Die Anwendung der vorgestellten performanten Maße auf vergleichende Anglizismen ist nun recht einfach: Wieder werden Wortpaare gebildet aus dem englischen Begriff und seiner eingedeutschten Entsprechung. Das Maß für das englische Fokuswort wird dann mittels englischem Korpus, das deutsche Fokuswort entsprechend im Deutschen. Die These, dass Anglizismen im Deutschen eine höhere semantische Spezifität aufweisen, lässt sich also experimentell bestätigen, wenn die Maße des englischen Fokusworts in der Regel höher als die das Deutschen ausfallen.

## 4.1 Textgrundlage

Um eine vergleichbare Textgrundlage im Deutschen wie Englischen zu verwenden, habe ich im Englischen auf den Reuters Corpus (Lewis et al., 2004) zurückgegriffen. Der Korpus umfasst in der verwendeten Version etwa 11.000 Nachrichtenartikel aus den späten 90ern. Im Deutschen habe ich den Tiger Korpus (Brants et al., 2004) verwendet, der eine vergleichbare Textmenge von Artikeln aus der Frankfurter Rundschau enthält.

## 4.2 Resultate

# 5 Konklusion

# 6 Caveats und Ausblick

Die vorgestellten Maße beschränken sich weitesgehend auf die reine statistische Verteilung. Es wäre bestimmt interessant, zusätzlich auch probabilistische Berechnungen zu verwenden..

Stillschweigend habe ich hier vorausgesetzt, dass die getesteten Maße auch in anderen Kontexten und Sprachen (etwa im Deutschen) greifen. Das habe ich zwar grob getestet,

aber keinem rigorosen Text unterzogen. Diese Hilfshypothesen erfordern eigentlich auch noch einiges an experimenteller Bestätigung.

Die hier vorgestellte Methode und das anschließende Experiment beinhalten einige vereinfachende Vorannahmen, die das Endergebnis negativ beeinflussen. Ein großer Faktor ist sicher, dass Wörter allein aufgrund ihrer lexikalischen Form identifiziert wurden. Es ist natürlich bekannt, dass gleiche lexikalische Formen unterschiedliche Bedeutungen tragen können, also ambig sind. Das hat direkten Einfluss auf den Featurevektor des Wortes - so werden auch alle Kookkurrenzen dazugezählt, die eigentlich einer anderen Wortbedeutung zugerechnet werden sollen. Abhilfe kann hier eine Word Sense Disambiguation schaffen, bei der entweder in der Präprozessierung oder bereits vorher, mittels eines annotierten Korpus, die verschiedenen Wortbedeutungen auseinandergehalten werden.

Auch wurde keinerlei unterscheidung gemacht hinsichtlich der Art des untersuchten Wortes und seiner Beziehung zu den umliegenden Wörtern, außer die direkte lexikalische Nachbarschaft. So wurden z.B. nicht Nomen untersucht und der Kontext aus Verben gebildet, die in direkter Subject-Prädikat-Relation stehen. Eine solche tiefergehende Analyse könnte u.U. dazu beitragen, die Erkennung von semantischer Spezifität zu präzisieren.

Die Auswertung der Resultate ist in der vorliegenden Form eher simplistisch. So wird einfach nur verglichen, welches Maß den höheren Wert berechnet hat und dann auf höhere / niedrigere semantische Sppezifität geschlossen. Was eigentlich gewünscht ist, wäre eine statistisch zuverlässige Aussage. Die Entscheidung über die Spezifität sollte mit einer gewissen Sicherheit erfolgen. Um das zu erreichen, müsste man zunächst über zufällig ausgewählte Samples aus dem Korpus eine generelle Verteilung der Differenzen der Spezifitätsmaße von Wortpaaren berechnen. Auf dieser Grundlage ließe sich die Standardabweichung der Differenz berechnen. Unter der Zusatzannahme, dass diese Differenz normalverteilt ist (was sich auch prüfen ließe), kann dann die statistische Signifikanz der Differenz der Maße eines Wortpaares berechnet werden, dessen semantische Spezifität man untersuchen will. So könnte dann etwa erst ab einem bestimmten Schwellenwert mit großer Sicherheit davon gesprochen werden, ob sich hier semantische Spezifität stark unterscheidet. Aber gleichzeitig wird auch die Präzision der Antwort erhöht, bei unzureichender Ausprägung enthält sich das Maß dann einer Wertung.

# 7 Anhang

## 7.1 Wordpaare

**food** beverage, dessert, bread, cheese, meat, dish, butter, cake, egg, candy, pastry, vegetable, fruit, sandwich, soup, pizza, salad, relish, olives, ketchup, cookie

**beverage** alcohol, cola

**alcohol** iquor, gin, rum, brandy, cognac, wine, champagner,

**meat** liver, ham

**dish** sandwich, soup, pizza, salad

**vegetable** tomato, mushroom, legume

**fruit** pineapple, apple, peaches, strawberry

**vehicle** truck, car, trailer, campers

**car** jeep, cab, coupe

**person** worker, writer, intellectual, professional, leader, editor, entertainer, engineer,technician, journalist, commentator, novelist

**intellectual** physicist, historian, chemist

**professional** physician, educator, nurse, dentist

**entity** organism, object

**animal** mammal, bird, dof, car, horse, chicken, duck, fish, turtle, snake

**mammal** cattle, dog, cat, horse

**bird** chicken, duck

**fish** herring, salmon, trout

**metal** alloy, steel, gold, silver, iron

**location** region, country, state, city

**substance** food, metal, carcinogen, fluid

**fluid** water

**commodity** clothing, appliance

**artifact** covering, paint, roof, curtain, decoration, drug

**publication** book, article

**fabric** wool, nylon, cotton

**facility** airport, headquarters, station

**structure** house, factory, store

**organ** heart, lung

## 7.2 Anglizismen

Airline, Babysitter, Bachelor, Bar, Basketball, Beach, Beat, Bestseller, Bits, Blackout, Blues, Bodybuilder, Boom, Boss, Box, Boys, Braindrain, Browser, Camper, Campus, Champion, clever, Coach, Cola, Comedy, Container, cool, Copyright, Date, Deal, Design, Drinks, Foul, Freak, Gameshow, Gangster, Hattrick, Hit, Hooligans, Image, Insider, Internet, Jazz, Keeper, Kids, Leasing, Lifestyle, Lobby, mail, Manager, Marketing, Meeting, model, Performance, Pixel, Poker, Pool, Punk, Quiz, Radar, Rapper, Scanner, Service, shop, Skateboard, Soccer, Sponsor, Stalker, Star, Start-up, Stewardess, Striptease, Surfer, super, Test, Training, Tricks, unfair, Website, Yacht, Yankee

# Literatur

**Brants, Sabine et al.:** TIGER: Linguistic Interpretation of a German Corpus. In: Research on Language and Computation, 2 2004, Nr. 4, 597–620 ⟨URL: http://dx.doi.org/10.1007/s11168-004-7431-3⟩, ISSN 1572–8706

**Caraballo, Sharon A. und Charniak, Eugene:** Determining the Specificity of Nouns From Text. In: In Proceedings SIGDAT-99. 1999, 63–70

**Han, Jiawei, Kamber, Micheline und Pei, Jian:** Data Mining: Concepts and Techniques. Morgan Kaufmann, 2011

**Heyer, Gerhard, Quasthoff, Uwe und Wittig, Thomas:** Text Mining: Wissensrohstoff Text – Konzepte, Algorithmen, Ergebnisse. 1. Auflage. W3L-Verlag, 2008

**Jones, Karen Sparck:** A Statistical Interpretation of Term Specificity and its Application in Retrieval. In: Journal of Documentation, 28 01 1972, Nr. 1, 11–21, ISSN 0022–0418

**Lewis, David D. et al.:** RCV1: A New Benchmark Collection for Text Categorization Research. In: J. Mach. Learn. Res. 5 Dezember 2004, 361–397 ⟨URL: http://dl.acm.org/citation.cfm?id=1005332.1005345⟩, ISSN 1532–4435

**Manning, Christopher D. und Schütze, Hinrich:** Foundations of Statistical Natural Language Processing. MIT Press, 1999

**Miller, G. und Leacock, Claudia:** Lexical Representations for Sentence Processing. In: **Ravin, Yael und Leacock, Claudia (Hrsg.):** Polysemy: Theoretical and Computational Approaches. Oxford University Press, 2000, 152–160

**Sahlgren, Magnus:** The Word-space model. Dissertation, University of Stockholm (Sweden), 2006

**Schütze, Hinrich und Pedersen, Jan O.:** A Cooccurrence-based Thesaurus and Two Applications to Information Retrieval. In: Intelligent Multimedia Information Retrieval Systems and Management - Volume 1. Paris, France, France: LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 1994, RIAO '94 ⟨URL: http://dl.acm.org/citation.cfm?id=2856823.2856847⟩, 266–274