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 8. September 2016)

Zusammenfassung. In dieser Arbeit versuche ich experimentell zu bestätigen, dass Angliszismen im Deutschen semantisch spezifischer sind als ihre englischen Ursprungswörter. Dazu entwickle ich zunächst verschiedene Verfahren, die statistische Eigenschaften der Wörter im Deutschen und Englischen untersuchen und Rückschlüsse auf die semantische Spezifität zulassen sollen. Zentral ist dabei die Idee, dass spezifischere Wörter einen dichteren Kontext haben, analog zu der Dichte eines Clusters im Wortraum. Die entwickelten Verfahren werden anschließend empirisch überprüft.

Anschließend wende ich diese Verfahren auf die zentrale These an. Dabei zeigen sich vielversprechende Resultate, die jedoch noch einige Unausgereiftheiten aufweisen. Daher wird die zentrale These nicht einheitlich experimentell bestätigt, die Ergebnisse indizieren jedoch das Potential der vorgestellten Maße.

Inhaltsverzeichnis

1 Einleitung

In dieser Arbeit verfolge ich zwei Ziele: Zum einen suche ich ein Maß für die semantische Spezifität von Worten. 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. Spezifischere Wörter treten seltener auf.
2. Spezifischere Wörter stehen mit weniger verschiedenen Wörtern im Kontext.
3. Spezifischere Wörter verfügen über 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/Schütze*, 1999, S. 4). Konkreter: Maße, die sich auf die Anzahl bestimmter Eigenschaften von Wörtern in Texten beziehen. Welche Berechnungen man anschließend mit den gezählten Werten anstellt, legt das mathematische Modell 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, S. 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 hat 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 ist ein Wort spezifischer, wenn es sich auf weniger Situationen anwenden lässt. Es schließt mehr Sachverhalte aus. Bezogen etwa auf die Tierwelt haben spezifischere Bezeichnungen eine geringere Extension: Mit 'Säugetier' kann man tausende verschiedene Tiere und Tierarten bezeichnen, mit 'europäischer Feldhamster' meine ich eine sehr spezielle Art von Tier.

Mit diesem intuitiven Verständnis von Spezifität wird auch schnell klar, wieso der Kontext des Wortes ein Indikator für dessen Spezifität sein kann. Ein spezifisches Fachwort tritt vorrangig zusammen mit anderen Wörtern aus dem selben Themenbereich auf, der Kontext ist eher eng gefasst. Ein allgemeines, vielseitig anwendbares Wort tritt mit allen möglichen Wörtern zusammen auf. Der Kontext ist sehr weit. Auch *Weeds/Weir/McCarthy* (2004) stellen fest, dass generelle Worte eher in einem weiten Kontext auftreten als Spezifischere.

Diese naheliegende Idee versuche ich vorligender Arbeit in ein mathematisches Modell zu übersetzen.

Semantische Spezifität und Anglizismen

Ein Anglizismus ist ein Wort im Deutschen, das einen Ursprung im Englischen hat. *Schütte* (1996, S. 38) definiert so:

> Ein Anglizismus ist ein sprachliches Zeichen, das ganz oder teilweise aus englischen Morphemen besteht, unabhängig davon, ob es mit einer im englischen Sprachgebrauch üblichen Bedeutung verbunden ist oder nicht.

Burmasova (2010, S. 216) fasst den Begriff etwas weiter:

> Alle sprachlichen Zeichen, deren englische Herkunft an der Form oder Semantik zu erkennen ist, gehören zu den Anglizismen.

Weiter werden in der Literatur über verschiedene Paradigmen Kategorisierungsvorschläge gemacht, um den Begriff Anglizismus genauer zu bestimmen. Für meine Zwecke sind die Definitionen oben aber bereits ausreichend: Worte im Deutschen, die eine erkennbare englische Herkunft haben. Um Deutsch und Englisch miteinander zu vergleichen, muss allerdings eine eine-zu-eins-Zuordenbarkeit bestehen.

Wie verhalten sich nun Anglizismen und ihre englischen Ursprungswörter in Bezug auf ihre semantische Spezifität? Findet eine Bedeutungsverschiebung statt? Ich stelle in dieser Arbeit die These auf, dass Anglizismen in ihrer Verwendung im Deutschen dazu tendieren, **spezieller** zu sein als ihre englischen Ursprungswörter.

Für diese These argumentiere ich hier nicht traditionell sprachwissenschaftlich oder versuche sie zu erklären, sondern versuche lediglich, gemäß der statistischen Eigenschaften zu bestätigen, dass es sich tatsächlich so verhält.

2 Ein Maß für die semantische Spezifität

Das Maß für die semantische Spezifität soll in der Lage sein, von zwei Wörtern das semantisch Spezifischere herauszusuchen.

Für die Auswahl eines passendes *Word Space Models* gibt es verschiedene Optionen. Im Information Retrieval weit verbreitet 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/Schütze*, 1999, S.296). Ebenfalls weit verbreitet 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 vorrangig auf Methoden aus dem Wortraum stützen. *Schütze/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 hin 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 Anzahl, mit der ein Wort im Korpus auftaucht, das ein Indiz für die Spezifität darstellen soll. *Caraballo/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: Der Unterbegriff ist 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 effizient. 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 der normierte **Document Frequency Score**

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

Weist ein Wort einen kleineren dfs-Wert auf, ist das nach *Caraballo/Charniak* (1999) eine gute Heuristik, 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, einen alternativen Überblick liefern *Turney/Pantel* (2010). 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: Das abstrakte Konzept der Bedeutungsähnlichkeit wird durch simple räumliche Nähe repräsentiert. Die statistischen Eigenschaften eines Wortes scheinen nach der Hypothese also auf nicht näher bestimmte Weise 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 Informationen fest, die sich mit Hilfe des WSM gewinnen 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.

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 wird dabei für zwei Wörter die Anzahl der Sätze, in denen die Wörter gemeinsam auftreten. Den Satzkontext des Fokuswortes ist 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

[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 Bedeutungsähnlichkeit.

betrachtet, also 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.
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 und (ii) welches Maß zur Berechnung der Kookkurrenz verwendet wird. Die verschiedenen Maße werden in Abschnitt 2.5 vorgestellt.

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 Zeile. Die Matrix ist dabei zwingend quadratisch, aber nicht unbedingt symmetrisch, z.B. im Falle der gerichteten Fensterkookkurrenzen.

Eine grundlegende These dieser Arbeit ist, dass 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, zu zählen, 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**

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

Ein kleinerer *nzds*-Wert eines Wortes spricht für eine höhere semantische Spezifität.

[2] Stopwords wurden bereits entfernt.

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. Daher ist bei jedem Maß zu beachten: $score_{ij}$ ist nicht zwingend gleich $score_{ji}$.

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 ist also $bin_{ij} = 1$ genau dann, wenn w_j in mindestens m Kontextfenstern von w_i auftritt.

2.5.2 Häufigkeit der Kookkurrenz

Die Häufigkeit der Kookkurrenz zählt jedes Auftreten der Kookkurrenz.

Sei K_{ij} die Menge aller Kontextfenster für Wort w_i, in denen das Wort w_j auftritt. Dann ist die **Häufigkeit**

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

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/Schütze* (1999, S.299), *Heyer/Quasthoff/Wittig* (2008, S. 213)).

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 Kontingenztabelle 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

Seien T die Kontingenztabelle zweier Wörter w_i und w_j, $O_{k,l}$ der Wert der Kontingenztabelle 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/Schütze* (1999, S. 169ff.).

2.6 Distanzmaße im Vektorraum

Um die Unterschiede der Featurevektoren im *Word Space Model* zu berechnen, habe ich wenige unterschiedliche Distanzmaße verwendet. In der Literatur finden sich einige, einen Überblick liefert *Cha* (2007).

2.6.1 Standardisierte 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 einer Dimension, so hat diese Dimension eine unverhältnismäßig große Auswirkung auf die Gesamtdistanz. Um diesen Effekt zu mindern, kann das Maß standardisiert werden. Dazu wird mit der reziproken Varianz des jeweiligen Features multipliziert. Die Varianz V_k ist dabei die quadrierte Standardabweichung jedes $k-ten$ Features aller Vektoren des Raumes.

Seien u und v die Featurevektoren der Worte w_i und w_j, u_k der k-te Wert des Featurevektors u und V_k die Varianz des Features k.
Dann ist die **Standardisierte Euklidische Distanz**

$$se(u,v) = \sqrt{\sum_k \frac{1}{V_k}(u_k - v_k)^2}.$$

2.6.2 Kosinusdistanz

Die Kosinusdistanz liefert ein bereits normalisiertes Maß dafür, wie stark zwei Vektoren korrellieren (*Manning/Schütze*, 1999, S. 300). Sie verhält sich analog zur euklidischen Distanz für normalisierte Vektoren. Zu beachten ist, dass oben die standardisierte, 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 Spezifität 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 sich ihre Bedeutungen. Welches der Wörter ist aber nun das semantisch spezifischere?

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

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 gibt.

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 (*Dunn* (1974), *Halkidi/Batistakis/Vazirgiannis* (2001)). 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

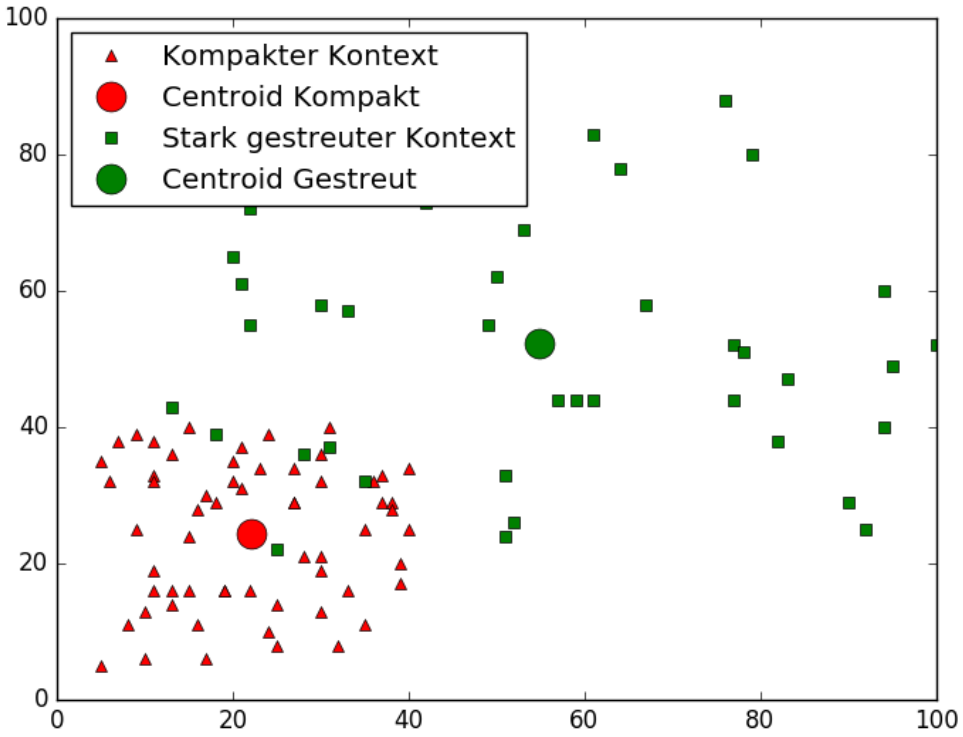

Abbildung 1: Streuung und Kompaktheit verschiedener Kontexte. Die beliebig dimensionalen Featurevektoren werden hier schematisch als Punkte im zweidimensionalen Raum dargestellt.

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 (vlg. Abbildung 1). Um das Maß nicht allzusehr zu verkomplizieren, habe ich zur Bestimmung der Verteilung der Distanz der einzelnen Featurevektoren zum Schwerpunkt den Durchschnitt 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**

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

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

Ich verwende für dieses Experiment den Brown Corpus (*Francis/Kucera*, 1979), der eine repräsentative Momentaufnahme des englischen geschriebenen Sprachgebrauchs aus den 1960er Jahren mit etwa einer Millionen Token darstellt.

Die Wörter werden dabei allein durch lexikalische Erscheinungsform analysiert, es findet der Einfachheit halber *keine* Bigram-Analyse statt. Die Wörter werden dabei mittels Porter-Stemmer Algorithmus (*Porter*, 1997) auf ihre Grundform reduziert. Anschließend werden die Wörter über diese lexikalische Grundform identifiziert.

Zusätzlich schließe ich Wörter von der Analyse aus, die weniger als zehn Mal im Korpus auftreten, um zumindest etwas die Aussagekraft der statistischen Eigenschaften sicher zu stellen.

3.2 Getestete Maße

Im Experiment kombiniere ich die vorgestellten Maße miteinander. Zur Referenz liste ich 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:

- standardisierte 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. Die

Kontext	K.-Maß	Score	Distanzmaß	Slug	Präzision
-	-	dfs	-	df	0.79
Satz	-	nzds	-	sent_nzds	0.77
Satz	Dice	mdcs	Kosinus	sent_dice_mdcs_cosi	0.74
Satz	Dice	mdcs	Euklidisch	sent_dice_mdcs_eucl	0.79
Satz	Frequenz	mdcs	Kosinus	sent_dice_mdcs_cosi	0.45
Satz	Frequenz	mdcs	Euklidisch	sent_dice_mdcs_eucl	0.22
Fenster	-	nzds	-	win_nzds	0.80
Fenster	Binär	mdcs	Kosinus	win_bin_mdcs_cosi	**0.81**
Fenster	Binär	mdcs	S-Euklidisch	win_bin_mdcs_seuc	0.79
Fenster	Frequenz	mdcs	Kosinus	win_freq_mdcs_cosi	**0.83**
Fenster	Frequenz	mdcs	S-Euklidisch	win_freq_mdcs_seuc	0.80
Fenster	Frequenz	scaled mdcs	S-Euklidisch	win_freq_sca_mdcs_seuc	0.79
Fenster	Dice	mdcs	Kosinus	win_dice_mdcs	**0.81**
Fenster	Chi Square	mdcs	Kosinus	win_chi_mdcs	**0.82**

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

Die verschiedenen Maße berechne ich nun zu jedem Wort der Wortpaare und vergleiche anschließend, welchen Begriff das Maß als semantisch spezifischer einstuft. Entspricht die Einstufung der Vorannahme, das der Unterbegriff semantisch spezifischer ist, wird das als richtige Einstufung gewertet. Die Präzision eines Maßes ist dann einfach der Anteil an geprüften Wortpaaren, den das Maß gemäß der 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/Leacock* (2000), die betonen, dass unser Verständnis von adequaten Kontexten noch nicht perfekt ist und man nicht von vornherein ausschließen sollte, dass sich Kontexte auch anders verstehen lassen. Ausgehend von diesem Ansatz stelle ich die Resultate als Funktion der Größe des Kontextfensters dar (s. Abbildungen 3, 4 und 5.) und gehe ohne theoretische Vorannahmen an diese Fragestellung heran.

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.

3.4.1 Satzkontext

Die Ergebnisse der Analyse nach Satzkontext ist ernüchternd (siehe Abbildung 2). Weder *nzds* noch *mcds* mit Kosinusmaß und standardisierter euklidischer Distanz erreichen die

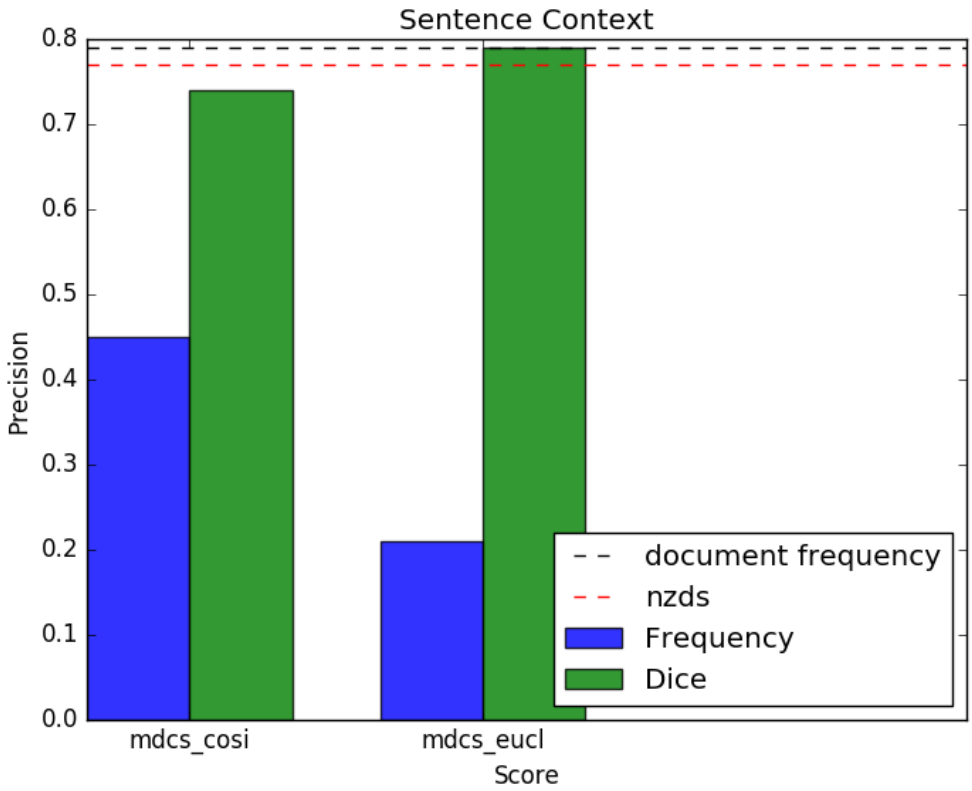

Abbildung 2: Präzision der Maße mit Frequenz und Dice-Koeffizient im Satzkontext.

Präzisionswerte des Vergleichswert *dfs*. 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 *dfs*.

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 Fenster den Vergleichswert.

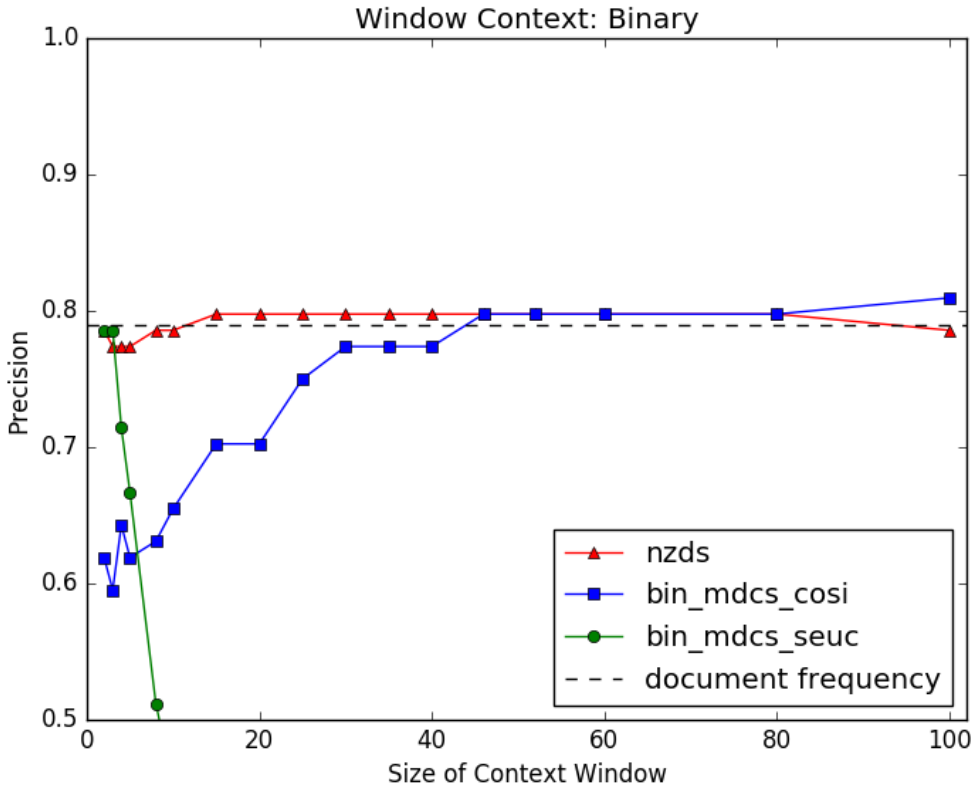

Abbildung 3: Präzision der Maße mit binärem Kookkurrenzmaß im Fensterkontext über Größe des Fensterkontextes.

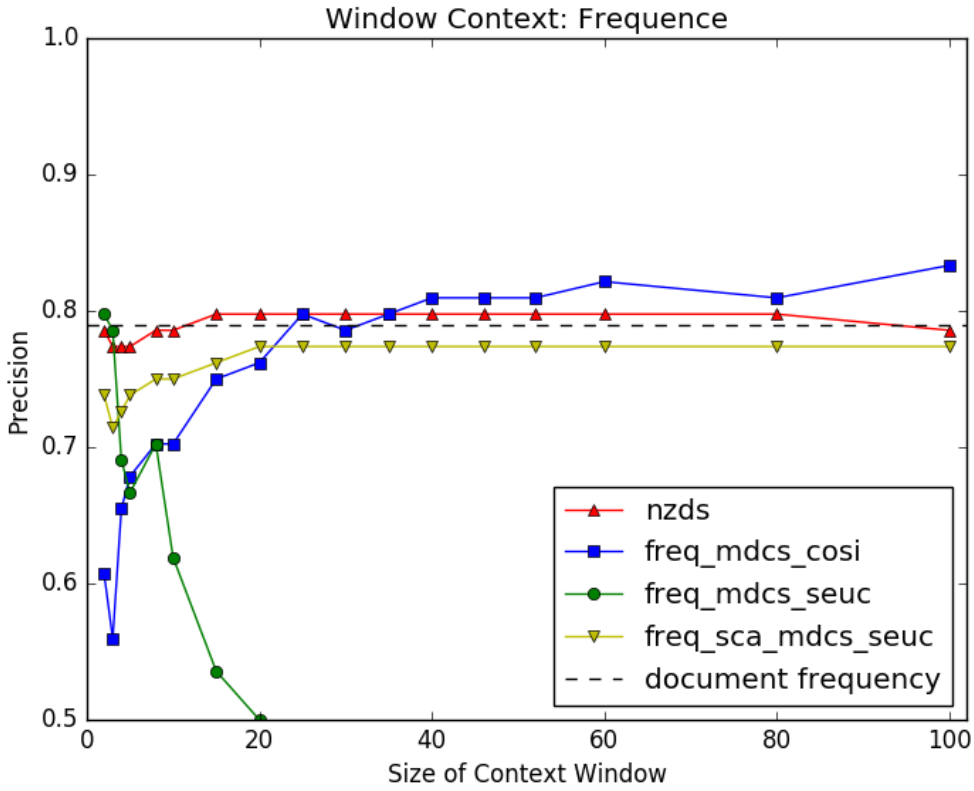

Abbildung 4: Präzision der Maße mit Frequency Kookkurrenzmaß im Fensterkontext über Größe des Fensterkontextes.

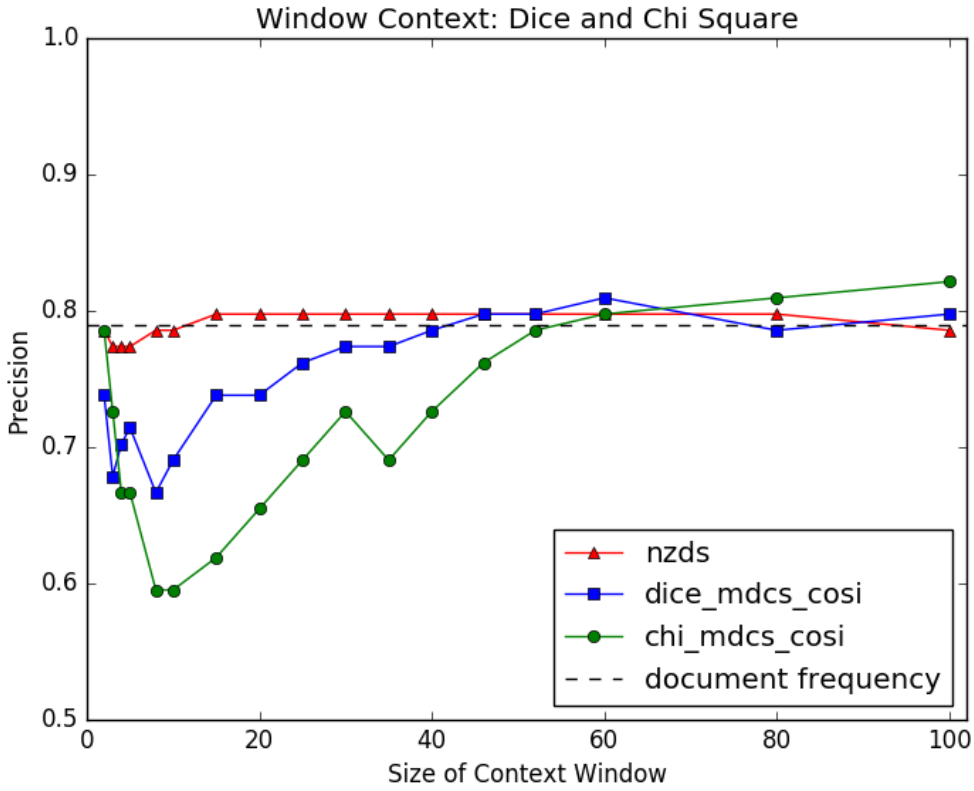

Abbildung 5: Präzision der Maße mit Dice bzw. Chi Square Kookkurrenzmaß im Fensterkontext über Größe des Fensterkontextes.

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 < 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 Präzision 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 der dfs. In der sehr starken Simplifizierung in den Vorannahmen sind noch einige Verbesserungsmöglichkeiten enthalten, ebenso in der Auswertung der Ergebnisse, mehr dazu in Kapitel 6. Interessant auch, dass sich die Ergebnisse mit *Sahlgren* (2006, S. 35) decken, der das Kosinusmaß für am besten geeignet hält. Alle hier vorgestellten Scores auf Basis einer Wortraumdistanz zeigen die höchste Präzision bei Verwendung des Kosinusmaßes.

4 Anwendung: Anglizismen

Die Anwendung der vorgestellten performanten Maße auf vergleichende Anglizismen ist nun recht einfach: Wieder werden Wortpaare gebildet aus dem englischen Wort und seiner deutschen Entsprechung. Das Maß für das englische Fokuswort wird dann mittels englischem Korpus berechnet, das deutsche Fokuswort entsprechend mittels Deutschem. 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 des 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.

Die verwendete Tokenisierung und das Stemming entspricht dem vorhergehenden Experiment. Die verwendete Liste von Anglizismen findet sich im Anhang. Von dieser Liste wurden nur solche verwendet, die öfter als zehnmal im Korpus auftraten. Die resultierende Liste findet sich im nächsten Abschitt.

4.2 Resultate

Tabellen 1-4 zeigen die Resultate der performantesten Maße aus dem vorhergehenden Experiment im Fensterkontext.

Da der vorliegende Tiger-Korpus keine Dokumentstruktur aufweist, wurde für dieses Experiment nicht der *dfs* berechnet. Der *nzds* hat jedoch zuverlässig sehr ähnlich oder besser performt und wird in diesem Experiment als Benchmark verwendet.

Die Tabelle geben jeweils an, ob für den jeweiligen Anglizismus die Hypothese bestätigt wird, dass das englische Wort allgemeiner verwendet wird als das Deutsche (True / False), d.h. ob der jeweilige Score des englischen Wortes höher ist als der des Deutschen.

Kookkurrenzmaß	-	binary	freq	chi_sq	dice
-	nzds	mdcs_cosi	mdcs_cosi	mdcs_cosi	mdcs_cosi
Bar	**True**	False	False	**True**	False
Bits	**True**	False	False	**True**	False
Boom	**True**	False	False	**False**	False
Date	**True**	False	False	**True**	False
Image	**True**	False	True	**False**	False
Manager	**True**	False	False	**True**	True
Marketing	**True**	False	False	**True**	True
Service	**True**	False	False	**True**	True
Star	**True**	False	False	**True**	False
Test	**True**	False	False	**True**	False

Tabelle 1: Fenstergröße 4

Kookkurrenzmaß	-	binary	freq	chi_sq	dice
-	nzds	mdcs_cosi	mdcs_cosi	mdcs_cosi	mdcs_cosi
Bar	**True**	False	False	**False**	False
Bits	**True**	False	False	**True**	False
Boom	**True**	False	False	**True**	False
Date	**True**	False	False	**True**	True
Image	**True**	False	False	**False**	False
Manager	**True**	False	False	**True**	True
Marketing	**True**	True	True	**True**	True
Service	**True**	False	False	**True**	True
Star	**True**	False	False	**True**	False
Test	**True**	False	False	**True**	True

Tabelle 2: Fenstergröße 25

4.3 Evaluation

Die Resultate dieses Experiments sind überraschend uneindeutig, gegeben die eher einheitlichen Ergebnisse aus dem vorangegangenen Experiment. Eigentlich würde man erwarten, dass sich die Scores relativ einig sind, was die höhere Spezifität der Worte angeht. Die Ergebnisse sind jedoch gemischt: Während der *ndzs* fast ausschließlich die Hypothese bestätigt, schwanken die anderen Maße sehr stark in der Ausprägung. Gegeben, dass die Maße für einen einheitlichen Korpus alle relativ zuverlässig sind, lässt das nun zwei verschiedene Erklärungen zu:

Kookkurrenzmaß	-	binary	freq	chi_sq	dice
-	nzds	mdcs_cosi	mdcs_cosi	mdcs_cosi	mdcs_cosi
Bar	**True**	False	False	**True**	False
Bits	**True**	False	False	**True**	False
Boom	**True**	False	False	**True**	False
Date	**True**	False	False	**True**	True
Image	**False**	False	False	**False**	False
Manager	**True**	False	False	**True**	True
Marketing	**True**	True	True	**True**	True
Service	**True**	False	False	**True**	True
Star	**True**	False	False	**True**	False
Test	**True**	False	False	**True**	True

Tabelle 3: Fentergröße 60

Kookkurrenzmaß	-	binary	freq	chi_sq	dice
-	nzds	mdcs_cosi	mdcs_cosi	mdcs_cosi	mdcs_cosi
Bar	**True**	False	False	**True**	False
Bits	**True**	False	False	**True**	False
Boom	**True**	False	False	**True**	False
Date	**True**	False	False	**True**	True
Image	**False**	False	False	**True**	False
Manager	**True**	False	False	**True**	True
Marketing	**True**	True	True	**True**	True
Service	**True**	True	False	**True**	True
Star	**True**	False	False	**True**	False
Test	**True**	False	False	**True**	True

Tabelle 4: Fentergröße 100

1. Die Hypothese ist falsch: Es ist nicht der Fall, dass Anglizismen dazu tendieren, semantisch spezifischer zu sein als ihr englisches Ursprungswort.
2. Die Vergleichsmaße über verschiedene Korpora sind nicht so unabhängig vom Korpus, wie ich erwartet habe.

Ich tendiere stark dazu, Erklärung 2 anzunehmen. Die Einheitlichkeit der Ergebnisse über verschiedene Korpora hinweg wurde noch nicht experimentell bestätigt. Das wäre definitiv wünschenswert, würde aber leider den Rahmen der vorliegenden Arbeit vollends sprengen. Beim Vergleich der Korpora fällt stark die unterschiedliche Dimensionierung auf: Englischer Korpus (Reuters): 4847 Wörter nach Normalisierung. Deutscher Korpus (Tiger): 6179 Wörter nach Normalisierung. Bei Reduzierung der Dimensionen auf 4951 durch Erhöhung des Frequency-Thresholds im Tigerkorpus auf 13 (statt 10) sind die erzielten Ergebnisse sehr viel eindeutiger und alle verwendeten Maße tendieren dazu, das englische Wort als genereller auszuzeichnen. Dies legt den Schluss nahe, dass die Hypothese zwar richtig ist, die verwendeten Maße jedoch nicht so unabhängig vom verwendeten Korpus sind wie gewünscht. Lediglich der *nzds* und *chi_mdcs_cosi* scheinen hier nicht so anfällig zu sein. Die Bestätigung der Robustheit der Maße steht also noch aus. Insbesondere wäre es interessant

zu sehen, *warum* sich diese Maße unterschiedlich verhalten. Für eine tiefergehende Analyse ist leider in dieser Arbeit kein Platz.

5 Konklusion

Im ersten Teil des Experiments konnte die von *Caraballo/Charniak* (1999) gefundene Performanz der *Document Frequency* bestätigt werden. Unter den neu formulierten Maßen wurden einige gefunden, die in einer festen Textmenge Wörter mit höherer Zuverlässigkeit als spezieller auszeichen können, als es die *Document Frequency* kann. Das kann man als kleinen Erfolg werten, der die aufgestellten Hypothesen (s. Einleitung) über den Zusammenhang von Spezifität und Kontext eines Wortes bestätigt.

Problematischer wird es beim Vergleich unterschiedlicher Korpora und unterschiedlicher Sprachen. Die experimentelle Bestätigung der Validität der Maße steht noch aus (s. nächster Abschnitt), allerdings besteht die berechtigte Hoffnung, das auch über verschiedene unangeglichene Korpora hinweg verläßlich das generellere Wort bestimmt werden kann. Insbesondere der *Non-Zero Dimensions Score* und der *Chi Square Mean Distance to Centroid Cosine Score* sind dabei vielversprechend.

Die These, dass Anglizismen im Deutschen semantisch spezifischer sind als ihre englischen Ursprungswörter, wurde in diesem Experiment also nicht eindeutig bestätigt, jedoch sehe ich Grund zur Annahme, dass durch Auswahl und Test angemessener Maße bessere Ergebnisse erzielt werden können.

6 Caveats und Ausblick

Die vorgestellt Auswertung ist vergleichsweise makroskopisch und soll einen Überblick über die Tauglichkeit verschiedener Verfahren liefern. Die Analyse einzelner Maße kann allerdings auch sehr aufschlussreich sein: So kann man sich etwa die Berechnungen einzelner Wortpaare genauer anschauen und so Schwachstellen oder Stärken des jeweiligen Maßes herausfinden.

Ein großer Vorteil der vorgestellten Maße ist, dass sie völlig generell und nicht themenbezogen sind, und zusätzlich ohne Trainingsdaten auskommen. Andere Verfahren, etwa über Klassifizierer, könnten allerdings performanter sein.

Ein Aspekt, der in dieser Arbeit nicht berücksichtigt wurde, ist die Gewichtung der Kookkurrenzen nach Distanz zum Fokuswort. Hier wurden alle Wörter, die im Kontextfenster auftauchen, gleich gewichtet, um die Maße nicht noch mehr zu verkomplizieren. Diese Änderung könnte bewirken, dass die signifikatnen Kookkurrenzen noch mehr im Vordergrund stehen.

Ich habe 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. Zwar habe ich auch in einem ersten Test überprüft, ob die Maße im Deutschen richtige Ergebnisse liefern (die Präzision der verschiedenen Maße äh-

nelt sich stark), eine experimentelle Bestätigung der Vergleichbarkeit der Maße aus einem Korpus mit denen eines anderen steht noch aus. Insbesondere verschiedene Verhältnisse der Korpora zueinander sollten hier Testparameter sein. Wie verhalten sich die Maße bei Korpora aus unterschiedlichen Themengebieten? Wie bei stark unterschiedlicher Größe?

Die hier vorgestellte Methode und das anschließende Experiment beinhalten einige vereinfachende Vorannahmen, die das Endergebnis negativ beeinflussen könnten. Ein großer Faktor ist wahrscheinlich, 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 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 Subjekt-Prädikat-Relation stehen. Eine solche tiefergehende Analyse könnte u.U. dazu beitragen, die Erkennung von semantischer Spezifität zu präzisieren.

Ein weiteres interessantes Maß wird in *Sahlgren/Karlgren* (2005) beschrieben. Dieses Maß ist dafür vorgesehen zu bestimmen, wie dicht (*dense*) ein Wortraum ist, d.h. wie kompakt oder gestreut der Wortraum mit Einträgen gefüllt ist. Die Ergebnisse werden dann dazu verwendet, verschiedene Korpora dahingehend zu vergleichen, wie thematisch ähnlich sie sich sind. Dazu werden zu jedem Punkt des Raumes die zehn nächsten Nachbarn bestimmt, dann wiederum die zehn nächsten Nachbarn. Der Durchschnitt der Anzahl der so gesammelten unique Worte bildet das Maß; es reicht von zehn (maximale Dichte) bis 100 (maximale Streuung). Dieser Ansatz scheint sehr gut auch auf semantische Spezifität zu passen und die Dichte eines Kontextes darstellen zu können. Um es noch unabhängiger vom verwendeten Korpus zu machen und so auch auf das Problem der Anglizismen anwenden zu können, kann man das Verhältnis vom Maß des Kontextes zum Maß des ganzen Korpus verwenden und hat so ein Maß von $0,1$ (maximale Dichte) bis $1,0$ (maximale Streuung). Ein interessanter Ansatz, der es leider nicht mehr in die Arbeit geschafft hat.

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 bzw. niedrigere semantische Spezifität 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 ge-

sprochen 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. Hat man auf diese Weise ein Maß für die Zuversicht in die Einschätzung des Maßes gefunden, lassen sich auch mehrere performante Scores aggregieren und so in einem Hybrid-Verfahren eine noch höhere Performanz erreichen. So kann bei widerstreitenden Ergebnissen zweier Maße dasjenige Maß mit der höheren Zuversicht entscheiden. Durch diese Aggregation entsteht dann ein neues, noch zuverlässigeres Verfahren.

Literatur

Brants, Sabine et al.: TIGER: Linguistic Interpretation of a German Corpus. In: Research on Language and Computation, 2 2004, Nr. 4, 597–620

Burmasova, S.: Empirische Untersuchung der Anglizismen im Deutschen: am Material der Zeitung Die Welt (Jahrgänge 1994 und 2004). University of Bamberg Press, 2010, Bamberger Beiträge zur Linguistik

Caraballo, Sharon A./Charniak, Eugene: Determining the Specificity of Nouns From Text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. 1999, 63–70

Cha, Sung-Hyuk: Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. In: International Journal of Mathematical Models and Methods in Applied Sciences, 1 2007, Nr. 4, 300–307

Dunn, J.C.: Well-Separated Clusters and Optimal Fuzzy Partitions. In: Journal of Cybernetics, 4 1974, Nr. 1, 95–104

Francis, W. Nelson/Kucera, Henry: The Brown Corpus: A Standard Corpus of Present-Day Edited American English. 1979, Brown University Liguistics Department

Halkidi, Maria/Batistakis, Yannis/Vazirgiannis, Michalis: On Clustering Validation Techniques. In: Journal of Intelligent Information Systems, 17 2001, Nr. 2, 107–145

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

Heyer, Gerhard/Quasthoff, Uwe/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 1972, Nr. 1, 11–21

Lewis, David D. et al.: RCV1: A New Benchmark Collection for Text Categorization Research. In: J. Mach. Learn. Res. 5 2004, 361–397

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

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

Porter, Martin F.: An Algorithm for Suffix Stripping. In: **Sparck Jones, Karen/Willett, Peter (Hrsg.):** Readings in Information Retrieval. San Francisco: Morgan Kaufmann Publishers Inc., 1997, 313–316

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

Sahlgren, Magnus/Karlgren, Jussi: Counting Lumps in Word Space: Density as a Measure of Corpus Homogeneity. In: **Consens, Mariano/Navarro, Gonzalo (Hrsg.):** String Processing and Information Retrieval: 12th International Conference. Proceedings. Berlin, Heidelberg: Springer, 2005, 151–154

Schütte, Dagmar: Das schöne Fremde: Anglo-amerikanische Einflüsse auf die Sprache der deutschen Zeitschriftenwerbung. VS Verlag für Sozialwissenschaften, 1996, Studien zur Kommunikationswissenschaft

Schütze, Hinrich/Pedersen, Jan O.: A Cooccurrence-based Thesaurus and Two Applications to Information Retrieval. In: Intelligent Multimedia Information Retrieval Systems and Management - Volume 1. Paris: Le Centre De Hautes Etudes Internationales D'Informatique Documentaire, 1994, 266–274

Turney, Peter D./Pantel, Patrick: From frequency to meaning : Vector space models of semantics. In: Journal of Artificial Intelligence Research, 2010, 141–188

Weeds, Julie/Weir, David/McCarthy, Diana: Characterising Measures of Lexical Distributional Similarity. In: Proceedings of Coling. 2004, 1015–1021

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

7.3 Verwendete Technologien

Das Experiment wurde komplett in `python 3.5.2` umgesetzt.

Den Zugang zu den Korpora habe ich mittels der Python-Bibliothek `nltk` realisiert. Auch die verwendeten Tokenisierer und Stemmer entstammen dieser Bibliothek. Die Berechnung der verschiedenen Kookkurrenzmaße erfolgten mit dieser Bibliothek.

Mathematische maschinennahe Berechnungen insbesondere für Matrizen habe ich mittels der Python-Bibliotheken `numpy` und `scipy` umgesetzt. Besonders bei der Berechnung der verschiedenen Distanzmaße waren diese Bibliotheken hilfreich.

Der Programmcode ist unter MIT License zugänglich unter
`https://github.com/conradfriedrich/termspec`.