

Linguistik
Computerlinguistik

Petra Steiner

Wortarten und Korpus

Automatische Wortartenklassifikation durch
distributionelle und quantitative Verfahren

Shaker Verlag
Aachen 2004

Bibliografische Information der Deutschen Bibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.ddb.de> abrufbar.

Zugl.: Münster (Westfalen), Univ., Diss., 2002

Copyright Shaker Verlag 2004

Alle Rechte, auch das des auszugsweisen Nachdruckes, der auszugsweisen oder vollständigen Wiedergabe, der Speicherung in Datenverarbeitungsanlagen und der Übersetzung, vorbehalten.

Printed in Germany.

ISBN 3-8322-2380-0

ISSN 1613-4532

Shaker Verlag GmbH • Postfach 101818 • 52018 Aachen

Telefon: 02407 / 95 96 - 0 • Telefax: 02407 / 95 96 - 9

Internet: www.shaker.de • eMail: info@shaker.de

Für meine Großmutter Agnes Julie Seifert geb. Bonk, meine Großmutter Herta Elise Steiner geb. Hasenbein und meine Großtante Gerda Klett geb. Seifert, die nicht studieren konnten, mich aber dennoch vieles gelehrt haben.

Vorwort	1
1 Einleitung	3
2 Die Kategorie <i>Wortart</i> in der Linguistik	5
2.1 Zum Wortbegriff	5
2.2 Anforderungen und Ziele einer Wortartenklassifikation	10
2.3 Zur Universalität von Wortarten	17
2.4 Kriterien für die Zuordnung von Wörtern zu Wortarten	24
2.4.1 Semantische Methoden zur Kategorisierung von Wortarten	25
2.4.2 Pragmatische Methoden zur Kategorisierung von Wortarten	29
2.4.3 Morphologische Kriterien zur Kategorisierung von Wortarten	29
2.4.4 Syntaktische und distributionelle Methoden zur Kategorisierung von Wortarten.....	31
3 Distributionelle Ansätze zur Wortartenkategorisierung.....	35
3.1 Die Begriffe der Kollokation und der Kolligation bei Firth.....	36
3.1.1 Der Begriff der Kollokation bei Firth und anderen	36
3.1.2 Der Begriff der Kolligation und die Bedeutung auf der grammatischen Ebene.....	38
3.2 Die Klassifikationsverfahren von Zellig S. Harris.....	39
3.2.1 Der Begriff der Distribution und der Umgebung	41
3.2.2 Klassifikation durch Substitution.....	42
3.2.3 Ermittlung komplementärer Segmente und Morpheme.....	44
3.2.4 Das Homonymenproblem.....	46
3.3 Kennedys distributionelle Wortartenbestimmungen für das Klassische Chinesisch ...	48
3.4 Das Wortartensystem von Bergenholtz und Schaefer.....	49
3.5 Clustering von Types mit dem Mutual Information-Index	50
3.6 Clustering von Types mit Log-Likelihood-Schätzungen.....	51
3.7 Reduktion von Tagsets bei Brants.....	52

3.8	Typeclustering als Input für die anschließende intellektuelle Wortartenklassifikation bei Brill und Marcus	52
3.9	Hypothesen zum kindlichen Spracherwerb und Type-Clustering mit Substitutionsverfahren.....	55
3.10	Clustern von Types bei Hughes und Atwell.....	56
3.11	Type- und Tokenclustering auf der Basis der Kotexte bei Schütze.....	58
3.12	Distributionelle Ansätze bei Rapp.....	59
3.13	Edit-Distanz und Kotext-Clustering bei Waterman.....	60
3.14	Resümee	62
4	Ergebnisse neuerer einzelsprachlicher Kategorisierungen	63
4.1	Typeklassifikationen auf der Grundlage von Cluster-Verfahren.....	64
4.1.1	Die Kategorisierungen.....	64
4.1.2	Beurteilung hinsichtlich der postulierten Anforderungen.....	67
4.2	Tokenklassifikationen auf der Grundlage von Cluster-Verfahren bei Schütze	70
4.2.1	Die Kategorisierungen.....	70
4.2.2	Beurteilung hinsichtlich der postulierten Anforderungen.....	70
4.3	Die Wortartenklassifikation von Bergenholtz und Schaefer	71
4.3.1	Die Kategorisierungen.....	71
4.3.2	Beurteilung hinsichtlich der postulierten Anforderungen.....	73
4.4	Die Wortartenklassifikation von Rapp auf der Grundlage von Bergenholtz und Schaefer.....	74
4.4.1	Die Kategorisierungen.....	74
4.4.2	Beurteilung hinsichtlich der postulierten Anforderungen.....	74
4.5	Die Kategorisierung des Brown-Korpus und des LOB-Korpus	75
4.5.1	Die Kategorisierungen.....	75
4.5.2	Beurteilung hinsichtlich der postulierten Anforderungen.....	76
4.6	Die Kategorisierung der Penn Treebank	77
4.6.1	Die Kategorisierungen.....	77
4.6.2	Beurteilung hinsichtlich der postulierten Anforderungen.....	77

4.7	Die Stuttgart-Tübinger Tagsets	78
4.7.1	Die Kategorisierungen.....	78
4.7.2	Beurteilung hinsichtlich der postulierten Anforderungen.....	79
4.8	Die Münsteraner Tagsets für das deutsche Münsteraner Korpus	80
4.8.1	Die Kategorisierungen.....	80
4.8.2	Beurteilung hinsichtlich der postulierten Anforderungen.....	80
4.9	Resümee	81
5	Statistische Eigenschaften von Wortartenkategorisierungen	83
5.1	Häufigkeitsklassenverteilungen und das „Zipsche Gesetz“	83
5.2	Der Zusammenhang zwischen der Frequenz und dem Informationsgehalt von Wörtern	86
5.3	Der Zusammenhang der semantischen und der grammatischen Mehrdeutigkeit mit der Frequenz von Wörtern	88
5.4	Modelle der Wortartenverteilungen	95
6	Die automatische Generierung von Wortartenklassen mit distributionellen, korpusbasierten Verfahren	105
6.1	Korpora als Grundlage von Kategorisierungsverfahren	105
6.1.1	Zur Definition des Korpusbegriffs.....	105
6.1.2	Zum Problem der Untersuchung eines opportunistischen Korpus.....	107
6.2	Die Verfahren	109
6.2.1	Die Operationalisierung des Tokens	109
6.2.2	Die Definition von Kotext	112
6.2.3	Kotextclustering-Verfahren	113
6.2.4	Verfahren zur Identifikation und Wortartenklassifikation von Funktionswörtern	128
6.2.5	Verfahren zur Identifikation und Wortartenklassifikation von Inhaltswörtern	137
6.2.6	Inkrementelle Zuweisung von Distributionsklassen (Kolligationsebene).....	139
6.2.7	Die Kombination der Verfahren	139
6.3	Implementierung	139
6.4	Evaluationsverfahren	141

6.5	Probleme bei der Wortartenkategorisierung im Deutschen	146
6.5.1	Die Unterscheidung von Artikeln, Relativpronomina und Demonstrativpronomina .	146
6.5.2	Das Problem der Abgrenzung zwischen Modal- und Vollverben.....	147
6.5.3	Das Problem der Abgrenzung zwischen Appellativa und Eigennamen sowie zwischen Monatsnamen, Jahreszahlen und Eigennamen	150
6.5.4	Das Problem der Abgrenzung zwischen Adverbien und prädikativen Adjektiven	155
6.5.5	Das Problem der Abgrenzung zwischen Adverbien und koordinierenden Konjunktionen	158
6.5.6	Das Problem der Abgrenzung zwischen Adverbien und pronominalen Indefinitpronomina (<i>mehr, wenig</i>).....	160
6.5.7	Das Problem der Abgrenzung zwischen attributiven Adjektiven, Zahlwörtern und attributiven Indefinitpronomina	162
6.5.8	<i>ein</i> : Adjektiv, attributives Indefinitpronomen, Artikel oder gar ein Nomen?	166
6.5.9	Das Problem der Abgrenzung von subordinierenden Konjunktionen und Interrogativadverbien	168
6.5.10	Ausrufe- und Fragezeichen mit ikonischem Charakter innerhalb eines Satzes.....	169
6.5.11	Die so genannten geschlossenen Wortklassen	169
6.6	Evaluation der Ergebnisse	172
6.6.1	Ergebnisse des Kotext-Clusterings über dem MKD und Auswahl der clusteranalytischen Verfahren und der Distanzmaße für das Kotextclustering.....	172
6.6.2	Die Ergebnisse auf der Grundlage der Testkorpora	197
7	Zusammenfassung der Ergebnisse und Beurteilung hinsichtlich der postulierten Anforderungen.....	223
8	Ausblick auf eine prototypische distributionelle Wortartenklassifikation.....	227
9	Literatur.....	229
10	Anhang: Die Tagsets für das deutsche Münsteraner Korpus	247

Verzeichnis der Abbildungen

Abbildung 1: Word Similarities from the Brown Corpus	65
Abbildung 2: Der Zusammenhang zwischen Häufigkeit und Informationsgehalt.....	87
Abbildung 3: Der Zusammenhang zwischen Häufigkeit, Polylexie und Lexeminventar	89
Abbildung 4: Anpassung des Types <i>groß/Groß</i> an die gemischte negative Binomialverteilung	93
Abbildung 5: Anpassung des Types <i>groß/Groß</i> an die negative Binomialverteilung.....	94
Abbildung 6: Der Kotext <i>eine#der</i> mit der Zwischenlänge 2 im MKD	113
Abbildung 7: Der Kotext <i>und#in</i> aus dem MKD mit Zwischenelementen der Länge 2	135
Abbildung 8: Der Kotext <i>aus#</i> , aus dem MKD mit Zwischenelementen der Länge 2	136
Abbildung 9: Teilbaum aus dem Ergebnis der geclusterten Kotexte mit der Quadrierten Euklidischen Distanz und dem Ward-Verfahren.....	182
Abbildung 10: Teilbaum aus dem Ergebnis der geclusterten Kotexte mit der Quadrierten Euklidischen Distanz und dem Ward-Verfahren (Ausschnitt aus Abbildung 9)	183
Abbildung 11: Teilbaum aus dem Ergebnis der geclusterten Kotexte auf der Basis standardisierter Kotextvektoren mit der Quadrierten Euklidischen Distanz und dem Ward-Verfahren.....	184
Abbildung 12: Teilbaum aus dem Ergebnis der geclusterten Kotexte auf der Basis standardisierter Kotextvektoren mit der Quadrierten Euklidischen Distanz und dem Ward-Verfahren (Ausschnitt aus Abbildung 11).....	185
Abbildung 13: Teilbaum aus dem Ergebnis der geclusterten Kotexte auf der Basis standardisierter Kotextvektoren mit der Maximum-Distanz und dem Single-Linkage-Verfahren	192
Abbildung 14: Ausschnitt aus dem mit Klassennamen versehenen MKD	201
Abbildung 15: Einige der Bestandteile der Zwischenelemente der Länge 2 des Kotexts <i>#das</i>	216
Abbildung 16: Einige Distributionsklassen mit zugehörigen Kotexten aus dem 13 Millionen-Korpus	217
Abbildung 17: Der Zusammenhang zwischen dem Singleton-Index und dem Anteil der Funktionswörter bei den Zwischenelementen eines Kotextes	220
Abbildung 18: Der Zusammenhang zwischen dem modifizierten Singleton-Index und dem Anteil der Funktionswörter bei den Zwischenelementen eines Kotextes.....	221

Verzeichnis der Tabellen

Tabelle 1: Häufigkeitsklassenverteilung des Birmingham Corpus	83
Tabelle 2: Häufigkeitsklassenverteilung beim Münsteraner Korpus Deutsch (MKD)	84
Tabelle 3: Verteilung der Zahl der Bedeutungen nach Lexemen	91
Tabelle 4: Verteilung der Zahl der möglichen Tags nach Tokens über das MKD	92
Tabelle 5: Die Wortartentags des Types <i>groß</i> im MKD.	92
Tabelle 6: Anpassung des Types <i>groß/Groß</i> an die gemischte negative und an die negative Binomialverteilung.....	93
Tabelle 7: Wortarten aus dem Wörterbuch amerikanischer Telefongespräche von French, Carter und Koenig (1930).	95
Tabelle 8: Die Häufigkeiten der (zusammengefassten) Wortartentags aus dem MKD	97
Tabelle 9: Ranghäufigkeitsverteilung der Wortarten in Bichsels <i>Und sie dürfen sagen,</i> <i>was sie wollen</i>	100
Tabelle 10: Ranghäufigkeitsverteilung der Wortarten in Brobrowskis <i>Betrachtung eines Bildes</i>	100
Tabelle 11: Die häufigsten Types im MKD	103
Tabelle 12: Die Häufigkeiten und relativen Häufigkeiten der häufigsten Character-Types der 2.623.220 Character-Tokens des MKD	110
Tabelle 13: Häufigkeitsverteilung über die Anzahl der Kotexte mit l, r der Länge 1 und $L(l, r) = 1$..	117
Tabelle 14: Die häufigsten Kotexte mit der Zwischenlänge 1 im MKD	118
Tabelle 15: Die häufigsten Kotexte mit der Zwischenlänge 1 des 13 Millionen-Korpus	118
Tabelle 16: Möglichkeiten der Übereinstimmung/Nichtübereinstimmung bei Vektoren mit binären Daten	127
Tabelle 17: Sich überlappende Cluster des MKD: Maximum-Distanzmaß, Single-Linkage-Verfahren über metrischen, nicht-standardisierten Daten	129
Tabelle 18: Sich überlappende Cluster: Maximum-Distanzmaß, Complete-Linkage-Verfahren über metrischen, nicht-standardisierten Daten	130
Tabelle 19: Sich überlappende Cluster: Quadrierte Euklidische Distanz, Zentroid-Verfahren über standardisierten Daten.....	131
Tabelle 20: 30 Klassen aus den jeweils ähnlichsten Clustern, Quadrierte Euklidische Distanz, Zentroid-Verfahren über standardisierten Daten.....	133
Tabelle 21: 30 Klassen aus den jeweils ähnlichsten Clustern, Quadrierte Euklidische Distanz, Complete-Linkage-Verfahren über metrischen, nicht-standardisierten Daten	134
Tabelle 22: Häufigkeitsverteilung über die Anzahl der Kotexte mit l, r der Länge 2 und $L(l, r) = 1$..	138
Tabelle 23: Kombinationen der verwendeten clusteranalytischen Verfahren und Distanzmaße mit der Form der Daten	173
Tabelle 24: Reduziertes Tagset für die Evaluation	174
Tabelle 25: Einige Kotexte mit der Zwischenlänge 1 aus dem MKD	179
Tabelle 26: Die Verteilungen über die Tag-Klassen bei den Kotexten aus Tabelle 25.....	180

Tabelle 27: Evaluation der einfachen Kotexte der Länge 1 über den 236 häufigsten Kotexten des MKD	180
Tabelle 28: Die ersten 25 Cluster beim Ward-Verfahren mit dem Distanzmaß der quadrierten Euklidischen Distanz über nicht-standardisierten Vektoren.....	186
Tabelle 29: Die ersten 30 Cluster beim Ward-Verfahren mit dem Distanzmaß der quadrierten Euklidischen Distanz über standardisierten Vektoren.....	187
Tabelle 30: Die Verteilungen über die Tag-Klassen bei den ersten 31 Clustern der nicht-standardisierten Daten (quadrierte Euklidische Distanz, Ward-Verfahren).....	188
Tabelle 31: Die Verteilungen über die Tag-Klassen bei den ersten 31 Clustern der standardisierten Daten (quadrierte Euklidische Distanz, Ward-Verfahren).....	189
Tabelle 32: Die einzelnen Kombinationen aus Distanzmaß, Clusteringverfahren und Daten (nichtstandardisiert/standardisiert) sortiert nach Durchschnitt, Median, Maximum der Precision	190
Tabelle 33: Die einzelnen Kombinationen aus Distanzmaß, Clusteringverfahren und Daten (nichtstandardisiert/standardisiert) sortiert nach Median, Durchschnitt, Maximum der Precision	191
Tabelle 34: Die einzelnen Kombinationen aus Distanzmaß, Clusteringverfahren und Daten (binär) sortiert nach Durchschnitt, Median, Maximum der Precision.....	193
Tabelle 35: Die Ergebnisse des Russell-Rao-Koeffizienten über binäre Daten sortiert nach Durchschnitt, Median, Maximum der Precision	194
Tabelle 36: Durchschnitt, Median, Minimum, Maximum, Standardabweichung der Precision von durch die K-means-Methode gebildeten 50 Clustern (metrische, nicht-standardisierte Daten)	194
Tabelle 37: Durchschnitt, Median, Extrema und Standardabweichung der Precision von 50 Clustern auf der Grundlage von hierarchischem Clustering (metrische, nicht-standardisierte Werte).....	195
Tabelle 38: Durchschnittswerte und gewichtete Durchschnittswerte von Precision, Recall und F-Maß der Wortartenklassen bei durch die K-means-Methode gebildeten 50 Clustern (metrische, nicht-standardisierte Daten)	195
Tabelle 39: Durchschnitts- und gewichtete Durchschnittswerte von Precision, Recall und F-Maß über die Tagklassen	195
Tabelle 40: Durchschnittswerte und gewichtete Durchschnittswerte von Precision, Recall und F-Maß der Wortartenklassen bei durch die K-means-Methode gebildeten 50 Clustern (metrische, standardisierte Daten)	196
Tabelle 41: Durchschnittswerte und gewichtete Durchschnittswerte von Precision, Recall und F-Maß der Wortartenklassen bei durch die K-means-Methode gebildeten 50 Clustern (binäre Daten)	196
Tabelle 42: Die Verteilungen über die Tag-Klassen im MKD bei einigen der 300 Cluster aus dem K-means-Verfahren nach 20 Iterationen	198

Tabelle 43: Evaluation der einfachen Kotexte der Länge 1 über den 13.730 häufigsten Kotexten des 13 Millionen-Korpus.....	199
Tabelle 44: 3 Cluster mit Artikeln über dem 13 Millionen-Korpus	200
Tabelle 45: Anteile der Artikel und Relativpronomina aus der Evaluation mit dem MKD.....	200
Tabelle 46: Kotexte mit l und r aus jeweils zwei Types bestehend mit Zwischenelementen aus dem MKD	211
Tabelle 47: Die Verteilungen über die Tag-Klassen bei den Kotexten	212
Tabelle 48: Kotexte mit l und r aus jeweils zwei Types bestehend mit Zwischenelementen aus dem 13 Millionen-Korpus	213
Tabelle 49: Durchschnittswerte und gewichtete Durchschnittswerte von Precision, Recall und F-Maß der Wortartenklassen bei durch die K-means-Methode gebildeten 100 Clustern aus den großen Kotexten des 13 Mio-Korpus.....	213
Tabelle 50: Cluster aus Kotexten mit $l, r = 2$ über dem 13 Millionen-Korpus mit Zwischenelementen und ihren Häufigkeiten.....	214
Tabelle 51: Singleton-Index, Häufigkeit, Kotext und Tokens mit Häufigkeiten für die zwei höchsten und die niedrigsten Singleton-Indizes	219
Tabelle 52: Klassen mit niedrigem Singleton-Index, aber hohem Anteil an Inhaltswörtern	222
Tabelle 53: Das große Münsteraner Tagset / Deutsch	250
Tabelle 54: Das kleine Münsteraner Tagset / Deutsch.....	252
Tabelle 55: Meta-Tags des Münsteraner Tagsets	252