

Projektbericht: Clustering von Adjektivklassen des Deutschen

Armin Schmidt
HS Maschinelles Lernen (Dr. Rainer Osswald), WS 08/09
Seminar für Computerlinguistik, Universität Heidelberg

15. April 2009

Zusammenfassung

In dieser Hausarbeit werden Experimente zum unüberwachten Lernen von Adjektivklassen des Deutschen vorgestellt. Die Clustering-Software Cluto wurde mit rein kontextuellen Merkmalen angewandt um automatisch möglichst sinnvolle Gruppierungen zu erzeugen. Um die Auswertung zu erleichtern, wurden maschinell annotierte Daten verwendet, deren Tagset attributive und prädikative Adjektive unterscheidet. Wie sich herausstellen wird, sind rein kontextuelle Merkmale nicht ausreichend, um hochwertige Gruppierungen herzustellen. Die Auswertung des Verfahrens erweist sich als insgesamt schwierig, da die meisten Adjektive sowohl attributiv als auch prädikativ verwendet werden können.

1 Einleitung und Literaturübersicht

Bisher war es vor Allem die traditionelle Semantikforschung, die sich mit den unterschiedlichen Gebrauchsformen und Kategorien von Adjektiven befasst hat. HAMMAN (1991) beispielsweise schlägt eine auf semantischen Eigenschaften basierte Unterteilung in prädikative und attributive Adjektive vor. Erstere bezeichnen demnach Eigenschaften ('der Ball ist rot'), letztere fungieren als Modifizierer ('ein roter Ball'). Eine gesonderte Stellung nehmen relationale Adjektive ein, welche einen auf eine bestimmte Eigenschaft bezogenen Vergleich ausdrücken ('der größere Ball'). Des Weiteren werden verschiedene Subklassen angenommen, beispielsweise die der absoluten Adjektive, die nicht steigerbar sind und typischerweise keine Antonyme besitzen ('griechisch', 'blau', 'quadratisch', 'verlobt').

Einige dieser Klassen lassen sich dabei anhand morphologischer Eigenheiten unterscheiden (attributive Adjektive werden flektiert, prädikative nicht), bei anderen ist die Zuordnung allein auf Ebene der lexikalischen Semantik möglich. Kontextuelle Merkmale kommen bei Hamman nicht zur Sprache. Eine allgemein bekannte und dennoch interessante Feststellung ist, dass bestimmte Adjektive im Deutschen entweder nur attributiv oder nur prädikativ verwendet werden ('ehemalig': '*der Präsident ist ehemalig'; 'allein': '*der alleine Junge').

Eine sehr umfangreiche Arbeit zur Typologie von Adjektivklassen ist die von DIXON/AIKHENVALD (2004), welche nicht nur die sprachübergreifenden Gemeinsamkeiten darstellt, sondern auch die Besonderheiten in vielen, typologisch äußerst unterschiedlichen, Sprachen untersucht. Adjektivkategorien werden hier nicht nur semantisch beschrieben, sondern auch in ihrem syntaktischen Verhalten.

Die maschinelle Klassifizierung von Adjektiven fand in der Computerlinguistik bisher wenig Beachtung. Eine Ausnahme ist die Arbeit von ALONSO/BOLEDA (2002), die sich mit dem unüberwachten Lernen von Adjektivklassen im Katalanischen befasst. Die vorliegende Hausarbeit orientiert sich grob an deren Arbeit in Bezug auf Zielsetzung und Auswahl der Methoden und Werkzeuge. Alonso und Boleda bedienen sich einer Reihe morphologischer sowie kontextueller Merkmale, die jeweils von einem Programm zur morphologischen Analyse bzw. einem Parser zur Analyse grammatischer Funktionen produziert werden. Ausgebaut wird die Arbeit in BOLEDA ET AL. (2005), in welcher kontextuelle und morphologische Merkmale gezielt einander gegenüber gestellt werden, dies allerdings nicht mehr im Rahmen eines unüberwachten Verfahrens. Die Autoren nutzen vielmehr die Anschaulichkeit von Entscheidungsbäumen für eine eher explorativ angelegte Fragestellung.

In der vorliegenden Hausarbeit soll untersucht werden, ob sich Adjektive im Deutschen allein anhand rein kontextueller Merkmale in die in der linguistischen Literatur beschriebenen Kategorien 'attributiv' und 'prädikativ' einteilen lassen. Dazu sollen unüberwachte Lernverfahren mithilfe der Clustering-Software Cluto (KARYPIS, 2003) eingesetzt werden.

Die Hausarbeit ist wie folgt aufgebaut: Abschnitt 2 diskutiert allgemeine Ansatzpunkte und Schwierigkeiten bei der Bearbeitung der Aufgabe. In Abschnitt 3 werden die Daten, deren Vorverarbeitung sowie die Berechnung der für das maschinelle Lernen notwendigen Merkmale dargestellt. Abschnitt 4 stellt die Experimente selbst vor und diskutiert deren Ergebnisse.

2 Herangehensweise

Eine der altbekannten Schwierigkeiten des Clustering ist die Auswertung maschinell gefundener Klassen. Solange keine Referenzklassifizierung der Daten zur Verfügung steht, bleibt oft nichts weiter übrig, als einzelne Cluster manuell zu inspizieren und auf ihre intuitive Erschließbarkeit hin zu untersuchen. Auch beim unüberwachten Lernen von Adjektivklassen wäre es interessant zu sehen, welche Gruppierungen ein Clustering-Verfahren wählen würde, ob diese mit traditionellen Kategorien der allgemeinen Linguistik übereinstimmen oder eventuell anderen, nichtsdestoweniger sinnvollen Einteilungen entsprechen. Problematisch ist bei solch einem Ansatz, dass die für das Clustering notwendigen Merkmale einerseits unbekannt sind und andererseits nicht direkt ausgewertet werden können, da klarerweise keine Referenzklassifizierung zur Verfügung stehen kann. Hinzu kommt, dass auch die Anzahl der zu erwartenden Klassen unbekannt ist und verschiedene Möglichkeiten durchprobiert werden müssten.¹

Um nicht auf manuelle Evaluierung angewiesen zu sein, habe ich mich in dieser Arbeit dazu entschlossen, auf bereits vorhandene Annotierung zurückzugreifen und die von der Clustering-Software Cluto (KARYPIS, 2003) produzierten Cluster anhand dieser automatisch zu bewerten. Als Datengrundlage diente der Dump der deutschen Wikipedia², der mit dem Tree-Tagger (SCHMID, 1994)³ wortartenannotiert wurde. Das vom Tree-Tagger für die Annotation deutscher Texte verwendete Stuttgart-

¹Einige Arbeiten versuchen, die Anzahl der Klassen automatisch zu bestimmen. Siehe CHOI (2000) für eine Anwendung im Bereich der Textsegmentierung.

²<http://de.wikipedia.org/wiki/Wikipedia:Download>

³Der Tree-Tagger kann unter <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/> heruntergeladen werden.

Tübingen-Tagset (STTS, SCHILLER ET AL., 1995) unterscheidet zwischen attributiv und nicht-attributiv verwendeten Adjektiven durch Zuweisung der Tags ADJA bzw. ADJD (siehe Abschnitt 1). Die so klassifizierten Daten sollen einerseits zur Extraktion der für das Clustering notwendigen Merkmale, als auch als Goldstandard für die spätere Evaluierung genutzt werden.

Wie bereits beschrieben, lassen sich aus rein linguistischer Sicht verschiedene Merkmale morphologischer und kontextueller Art bestimmen, anhand derer zwischen den Arten von Adjektiven unterschieden werden kann. Da im Rahmen dieser Hausarbeit keine morphologische Analyse zur Verfügung stand, habe ich mich auf rein kontextuelle Merkmale beschränkt. Nun ließen sich anhand der Beschreibung des Tagsets (SCHILLER ET AL., 1995) einige wenige solcher feststellen und so die Merkmalsextraktion auf diese beschränken. Zum Beispiel sollten nicht-attributive Adjektive insbesondere nach Verben oder Nomen auftreten, attributive jedoch vor allem vor Nomen. Um dem Lernverfahren jedoch die Möglichkeit zu geben, die Beziehungen zwischen Adjektiven und anderen Wortarten selbstständig zu bewerten und zusätzlich zu den typischen Fällen eventuell andere charakteristische Merkmale in Betracht ziehen zu können, habe ich die Merkmalsmatrix so aufgebaut, dass jedes Adjektiv durch die Wortart seines Vorgängers sowie Nachfolgers beschrieben wird. Da Cluto die Möglichkeit bietet, diejenigen Merkmale anzuzeigen, welche eine Klasse am besten beschreiben, kann in der Auswertung so auch überprüft werden, ob die ursprünglichen Beschreibungen und damit einhergehenden Hypothesen über die kontextuellen Charakteristika einer Wortklasse empirisch bestätigt werden können.

Schließlich stellen sich noch zwei Fragen: „Wie sollen die Merkmale dargestellt werden?“ und „Soll auf Oberflächenformen oder auf Lemmata gearbeitet werden?“ Zur Beantwortung der ersten Frage habe ich mich an der von ALONSO/BOLEDA (2002) beschriebenen Vorgehensweise orientiert und ein kontextuelles Merkmal so kodiert, dass es der Häufigkeit entspricht, mit welcher es mit einem Adjektiv vorkommt (zur genauen Berechnung siehe Abschnitt 3).

Die zweite Frage wirft ein paar grundlegende Fragen nach dem Ziel der Arbeit auf. Eine auf Lemmata basierende Merkmalsmatrix ist weniger stark gestreut und besitzt daher potentiell mehr Aussagekraft. Auf der anderen Seite können die allermeisten Adjektive sowohl attributiv als auch prädikativ gebraucht werden, so dass die Verwendung von deren Lemmata weniger gut zwischen den einzelnen Klassen unterscheiden könnte. Ein Clustering-Verfahren würde so eventuell eher feststellen, welche Gebrauchsart für ein bestimmtes Adjektiv typisch ist, nicht jedoch zu welcher Klasse es gehört.

Bei der Verwendung von Oberflächenformen, d. h. den flektierten Adjektivformen, ist dieses Problem weniger stark ausgeprägt, besteht jedoch weiterhin. Ein flektiertes Adjektiv wie 'schneller' kann sowohl attributiven ('ein schneller Wagen') als auch prädikativen ('der schneller fahrende Wagen') Gebrauch signalisieren und ohne weitere Information nicht disambiguiert werden.

In den hier vorgestellten Experimenten habe ich mich auf die lemmabasierte Untersuchung beschränkt.

3 Vorverarbeitung der Daten

Die POS-annotierten Daten wurden mit der Skriptsprache Python in das Eingabeformat für Cluto umgewandelt. Cluto akzeptiert zwei verschiedene Eingabeformate (*Sparse Matrix Format* und *Dense Matrix Format*). Da die für diese Aufgabe generierten Matrizen viele Nullstellen enthielten, habe ich das *Sparse Matrix Format* verwendet (ein Beispiel folgt weiter unten).

Das STTS beinhaltet insgesamt 54 Tags, von denen 49 als Merkmale in Frage kommen (unter Anderem habe ich Interjektionen und nicht-satzbeendende Interpunktionszeichen aufgrund mangelnder Relevanz unbeachtet gelassen). Die Eingabematrizen für den lemma- als auch den oberflächenbasierten Ansatz besteht dem entsprechend aus insgesamt 98 Spalten: 48 für das dem Adjektiv vorangehende sowie das ihm nachfolgende Tag. Jede Zeile entspricht einem 'Objekt', d.h. einem Lemma und enthält die entsprechenden Merkmalswerte in den ihnen entsprechenden Spalten.

Bei i verschiedenen Merkmalen soll ein Merkmalswert die Häufigkeit n_i ausdrücken, mit welchem ein Adjektiv mit einer bestimmten Wortart vorkommt im Verhältnis zu der Häufigkeit N mit welcher das Adjektiv insgesamt vorkommt:

$$\frac{n_i}{N} \quad (1)$$

Da die Werte hier sehr klein werden können und Python sehr kleine Brüche in der Cluto nicht bekannten Exponentialnotation (z.B. $7.5e-5$) darstellt, habe von 1 der Logarithmus berechnet:

$$\log \frac{n_i}{N} \quad (2)$$

Da es oft vorkommt, dass die Häufigkeit einer bestimmten vorangehenden oder nachfolgenden Wortart die Gesamthäufigkeit eines Adjektivs ausmacht, d. h. $n_i = N$, und da $\log 1 = 0$, Merkmalswerte von 0 jedoch vermieden werden sollten, wurde N um 1 erhöht. Zusätzlich wurde zur besseren Lesbarkeit der Merkmalswert mit -1 multipliziert, um einen positiven Wert zu erhalten:

$$-\log \frac{n_i}{N + 1} \quad (3)$$

Insgesamt wurden gut 26 Mio. Adjektive extrahiert, von denen knapp 20 Mio. das Tag ADJA und gut 6 Mio. das Tag ADJD zugewiesen hatten. Es wurden Merkmalsmatrizen sowohl für das lemma- als auch das oberflächenbasierte Clustering generiert und in Clutos Eingabeformat transformiert. Abbildung 1 zeigt zwei zufällige Zeilen der lemmabasierten Matrix samt den dazu gehörenden Adjektiven (Lemmata).⁴

```
(panikartiger) 14 0.741937344729 51 2.35137525716 61 3.04452243772
63 1.09861228867
```

```
(zusammengestellter) 1 2.19722457734 5 2.19722457734 14 1.50407739678
57 1.09861228867 59 2.19722457734
```

Abbildung 1: Zwei zufällige Zeilen der Eingabematrix für das lemmabasierte Clustering. Die Lemmata sind in der tatsächlichen Eingabedatei nicht vorhanden.

⁴Dass die Wörter *panikartiger* und *zusammengestellter* offensichtlich flektiert und keine Lemmata im linguistischen Sinne sind, zeigt die Schwächen automatischer Wortartenannotation und gleichzeitig eine der Fehlerquellen bei dieser Arbeit auf.

4 Experimente

4.1 Cluto

Cluto (KARYPIS, 2003)⁵ ist ein Software-Paket, das verschiedene Clustering-Algorithmen sowie Werkzeuge zur Evaluierung in Form von alleinstehenden Programmen als auch in Form einer Programm-bibliothek zur Verfügung stellt. Dabei werden sowohl partitionierende ('teilende'), agglomerative ('zusammenfassende'), als auch auf Graphen basierende (*graph-partitioning*) Algorithmen angeboten. Alle von Cluto implementierten Algorithmen betrachten eine Clustering-Aufgabe als ein Optimierungsproblem, bei dem versucht wird, eine bestimmte Zielfunktion (*criterion function*, auch: *objective function*) lokal oder global zu maximieren bzw., je nach Art der Funktion, zu minimieren. Die unterschiedlichen Algorithmen und Zielfunktionen werden in ZHAO ET AL. (2005); ZHAO/KARYPIS (2005) beschrieben. Das mit Cluto ausgelieferte Programm `vccluster` bietet ein Fülle von Funktionen, die über diverse Kommandozeilenoptionen gesteuert werden können und den Nutzer die Art des Clustering-Algorithmus und der Zielfunktion, verschiedene Datenmodifizierungen, Auswertungsmöglichkeiten sowie Ein- und Ausgabeformate festlegen lassen.

4.2 Programmparameter

Im Folgenden sollen die Optionen bei Durchführung des Clustering-Experiments mit Cluto beschrieben werden. Ein Beispielaufruf des Programms sieht so aus:

```
> vccluster -clmethod=rbr -crfun=h2 -sim=corr -colmodel=idf
-rlabelfile=rows_lemmaFeatMatrix.txt -rclassfile=lemmaClassLabels.txt -showfeatures
-showsummaries=itemsets lemmaFeatMatrix.txt 2
```

Die Bedeutungen der einzelnen Optionen schlüsselt sich wie folgt auf:

- `-clmethod`. Der Clustering-Algorithmus. Hier kann zwischen partitionierenden (`rb`, `rbr`, `direct`), agglomerativen (`agglo`, `bagglo`) und verschiedenen graphbasierten Algorithmen gewählt werden.
- `-crfun`. Die Zielfunktion. Es werden verschiedene Funktionen angeboten, welche entweder ein lokales oder globales Optimum in den Vordergrund stellen.
- `-sim`. Das Ähnlichkeitsmaß zur Berechnung der Relation zweier Objekte. Für partitionierende Algorithmen stehen hier das Kosinusmaß und der Korrelationskoeffizient zur Verfügung.
- `-colmodel`. Bei Wahl des Parameters `idf` wird in Anlehnung an die im Information Retrieval oft genutzte *Inverse Document Frequency* das Gewicht bei besonders vielen Objekten auftretender Spalten reduziert.
- `-rlabelfile`. Die Datei, welche die Bezeichnungen der einzelnen Objekte enthält. In meinem Experiment entsprechen diese den Lemmata bzw. Oberflächenformen.

⁵<http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview>

cid	Size	ISim	ISdev	ESim	ESdev	Entpy	Purty	ADJD	ADJA
0	13801	+0.255	+0.071	+0.123	+0.047	0.968	0.604	5462	8339
1	18307	+0.160	+0.052	+0.123	+0.062	0.530	0.880	2199	16108

Tabelle 1: Ergebnisse für das lemmabasierte Clustering

- `-rclassfile`. Die Datei, welche die tatsächlichen Klassenzugehörigkeiten enthält. Diese Information wird zur Auswertung des Clusterings, d. h. zur Berechnung der Reinheit einzelner Klassen verwendet.
- `-showfeatures`. Weist das Programm an, die Merkmale auszugeben, welche die einzelnen Klassen am besten beschreiben, d. h. am charakteristischsten für eine Klasse sind.
- Die letzten beiden Kommandozeilenparameter geben die Datei an, welche die Eingabematrix enthält sowie die Anzahl der gewünschten Klassen.

Während der Experimente habe ich verschiedene Algorithmen und Zielfunktionen ausprobiert und feststellen müssen, dass die Ergebnisse sich zwar in Einzelheiten unterscheiden, die Gesamtqualität der Cluster jedoch nur wenig beeinflussen.

Weiterhin bleibt festzustellen, dass bei einem Clustering in nur zwei Klassen, wie es in den hier durchgeführten Experimenten der Fall ist, viele der Unterschiede, die bei der Auswahl verschiedener Optionen bemerkbar sein können, kaum ins Gewicht fallen. Beispielsweise dürften die partitionierenden Algorithmen *rb*, *rbr* und *direct* bei nur zwei Klassen nahezu identische Ergebnisse liefern.

4.3 Auswertung

Die besten Ergebnisse für das lemmabasierte Clustering konnten mit einer Kombination des *Repeated-Bisectioning*-Algorithmus (*rb*), der Zielfunktion I_1 , dem Kosinusmaß und IDF-Normalisierung erreicht werden (Tabelle 1).

Die erste Spalte (*cid*) der Ergebnistabelle benennt die Cluster-ID. *Size* bezieht sich auf die Größe der Cluster. Hier fällt auf, dass das Cluster 1 etwas größer ist als Cluster 0. Tatsächlich kommen attributive Adjektive (ADJA) wesentlich häufiger als prädikative vor. In den von mir verwendeten Daten stehen, wie in Abschnitt 3 erwähnt, knapp 20 Mio. vornehmlich attributive gut 6 Mio. vornehmlich prädikativen Adjektiven bzw. deren Lemmata gegenüber. Die Spalten 3 bis 6 geben die Werte für die Cluster-interne bzw. -externe Ähnlichkeit und Standardabweichung an. Die Werte scheinen hier tendentiell richtig zu liegen, jedoch sollte gerade die Cluster-interne Ähnlichkeit sich im Idealfall wesentlich deutlicher von der Cluster-externen abheben. Die Spalten überschrieben mit *Entpy* bzw. *Purty* beziehen sich auf die Cluster-Entropie und die Cluster-Reinheit. Im Idealfall sollten die Entropiewerte hier möglichst klein, die Reinheitswerte jedoch hoch sein (beide haben den Wertebereich $[0, 1]$). Die Gesamtentropie beträgt 0.718, die Gesamtreinheit 0.761. Die letzten beiden Spalten enthalten die Anteile der tatsächlichen Klassen innerhalb der Cluster. Nimmt man als gegeben, dass Cluster 1

0	deskr.	\$. (5.3%)	_APPR (4.5%)	_VAFIN (4.5%)	_VVFIN (4.2%)	_VVPP (4.1%)
	diskr.	ART_ (11.1%)	APPR_ (5.4%)	\$_ (5.1%)	VAFIN_ (5.1%)	_VVFIN (4.7%)
1	deskr.	ART_ (11.4%)	APPR_ (9.0%)	_ADJA (7.4%)	ADJA_ (7.1%)	KON_ (6.6%)
	diskr.	ART_ (11.1%)	APPR_ (5.4%)	\$_ (5.1%)	_VAFIN (5.1%)	_VVFIN (4.7%)

Tabelle 2: Diskriminierende und deskriptive Merkmale. Ein Unterstrich ('_') markiert die Position eines Adjektivs.

(cid 1) der Klasse der attributiven Adjektive entspricht, lassen sich diese Werte zur Berechnung der Genauigkeit des Verfahrens nutzen. Die Akkuratheit des Clusterings beträgt demnach 0.67.

Obwohl dieser Wert gar nicht schlecht anmuten mag, sollte man sich nicht über die Tatsache hinwegtäuschen lassen, dass Cluster 0 mehr falsche als richtige Adjektive enthält. An dieser Stelle muss allerdings noch einmal auf die Schwächen der Auswertung hingewiesen werden: Wie bereits festgestellt wurde, können Adjektive mit wenigen Ausnahmen sowohl attributiv als auch prädikativ verwendet werden. Die Zuweisung eines Lemmas zu einer der beiden Klassen kann maximal eine Tendenz beschreiben, denn echte Klassenzugehörigkeit gibt es in den seltensten Fällen.

Die Merkmale, welche die einzelnen Cluster am besten beschreiben, sind in Tabelle 2 angegeben. Hierbei fällt auf, dass ihr jeweiliger Beitrag einigermaßen gering ausfällt (fast alle Prozentwerte liegen unter 10%). Des Weiteren ist bemerkenswert, welche Merkmale es sind, die herangezogen wurden. Die Position eines benachbarten Nomens spielt offenbar keine große Rollen. Dies kann allerdings auch ein ungewollter Nebeneffekt der Option `-colmodel=idf` sein, welche den Einfluss von Nomen reduziert haben mag.

5 Fazit

In dieser Hausarbeit habe ich Experimente zum unüberwachten Lernen von Adjektivklassen im Deutschen vorgestellt. Ich habe grundlegende Überlegungen zum Verfahren dargestellt, wichtige Punkte bei der Vorverarbeitung der Daten und den Experimenten selbst diskutiert und die Ergebnisse ausgewertet.

Insbesondere der Umstand, dass die allermeisten Adjektive sowohl attributiv als auch prädikativ verwendet werden können, erschwert die Auswertung sehr. Auch lässt sich abschließend feststellen, dass kontextuelle Merkmale, zumindest dann, wenn sie sich lediglich auf die Wortarten benachbarter Wörter beziehen, zur Beschreibung der Charakteristika einer Klasse nicht genügen. Eine Folgeuntersuchung sollte versuchen, morphologische und funktionale Merkmale einzubeziehen, wie dies in anderen Arbeiten bereits getan wurde (vgl. BOLEDA ET AL., 2005). Auch sollte größeres Augenmerk auf klassenreine Adjektive gelegt werden, d.h. solche, die ausschließlich dem Paradigma einer, nicht jedoch dem der anderen Klasse angehören können. Die Verteilung dieser könnte mehr Aufschluss über die Richtigkeit der verwendeten Merkmale geben.

Literatur

- ALONSO, L./BOLEDA, G. (2002): An Approach to Catalan Adjective Lexical Classes by Clustering, Workshop: Quantitative Investigations in Theoretical Linguistics, Universität Osnabrück, <http://www.cogsci.uni-osnabrueck.de/qitl/QITL1/QITL-Dateien/Abstracts/alonsoboleda.htm>.
- BOLEDA, G./BADIA, T./IM WALDE, S. (2005): Morphology vs. Syntax in Adjective Class Acquisition, in: ACL-SIGLEX 2005, 100, S. 77.
- CHOI, F. (2000): Advances in Domain Independent Linear Text Segmentation, in: Proceedings of the first conference on North American chapter of the Association for Computational Linguistics, S. 26–33.
- DIXON, R./AIKHENVALD, A. (2004): Adjective Classes. A Cross-linguistic Typology, Oxford University Press, USA.
- HAMMAN, C. (1991): Semantik. Ein Internationales Handbuch der Zeitgenössischen Forschung, Kap. Adjektivsemantik, S. 657–673, de Gruyter.
- KARYPIS, G. (2003): CLUTO. A Clustering Toolkit, University of Minnesota, Dep. of Computer Science, Mineapolis, MN55455, 2. Aufl.
- SCHILLER, A./TEUFEL, S./STÖCKERT, C./THIELEN, C. (1995): Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS.
- SCHMID, H. (1994): Probabilistic part-of-speech tagging using decision trees, in: Proceedings of International Conference on New Methods in Language Processing, Bd. 12, Manchester, UK.
- ZHAO, Y./KARYPIS, G. (2005): Criterion functions for document clustering, University of Minnesota.
- ZHAO, Y./KARYPIS, G./FAYYAD, U. (2005): Hierarchical clustering algorithms for document datasets, in: Data Mining and Knowledge Discovery, 10(2), S. 141–168.