

# Kollokationen und semantisches Clustering

Martin LÄUTER    Uwe QUASTHOFF

## Einleitung

Kollokationen bilden eine gute Möglichkeit, das semantische Umfeld eines Wortes zu erfassen. Dabei werden, ausgehend von einem Wort, alle die Wörter zusammengestellt, die in einem Korpus signifikant häufig in der Nähe dieses Wortes auftreten. Von den Autoren werden verschiedene Signifikanzmaße und Nachbarschaftsbegriffe (sowie natürlich auch unterschiedliche Korpora) verwendet, die die Ergebnisse nicht direkt vergleichbar machen. Bei einer Bewertung der Ergebnisse verschiedener Verfahren werden in der Regel die (entsprechend dem jeweiligen Modell) stärksten Kollokationen durch eine intellektuelle Bewertung auf semantische Nähe zum Ausgangswort hin untersucht.

Diese semantische Nähe drückt sich häufig durch eine bestimmte Relation aus, in der sich die beiden Wörter befinden. Von speziellem Interesse ist es, auch diese Relationen automatisch zu bestimmen, d.h. das Verfahren zur Kollokationsbestimmung so zu modifizieren, dass nur Kollokationen mit bestimmten Relationen gefunden werden.

Ziel dieser Arbeit ist es, mit automatischen Mitteln mehrere Wörter in einer Kollokationsmenge anzugeben, deren semantischer Zusammenhang klar ist. Dies scheint für die folgenden drei Gruppen möglich zu sein:

- Gruppen von Kohyponymen, d. h. Unterbegriffe zu einem gemeinsamen Oberbegriff;
- Feste Wendungen;
- Mehrwort-Eigennamen.

Das verwendete Verfahren startet mit signifikanten Paaren und bildet daraus immer größere Mengen von Wörtern, die signifikant häufig gemeinsam auftreten. Dabei kann unter bestimmten Bedingungen an die Anzahlen solcher Obermengen eine Aussage über die Relation getroffen werden.

## Das Verfahren

Ausgangspunkt sind Kollokationspaare (d. h. Paare von Wortformen, die statistisch signifikant häufig gemeinsam in einem Satz vorkommen), die mit einem Verfahren beruhend auf einem G-Test für eine angenommene Poisson-Verteilung ermittelt wurden (WEBER 1980). Das Herangehen ist ausführlich in QUASTHOFF (1999b) dargestellt. Dieses Verfahren ist statistisch zuverlässig und lässt sich einfach von Paaren auf Tripel und größere Tupel verallgemeinern. Im Fall von Kollokationspaaren liefern die Daten gute Übereinstimmung mit Daten, die mit der log-Likelihood-Methode ermittelt wurden (DUNNING 1994).

Aus ca. 7,5 Millionen deutschsprachigen Beispielsätzen wurden 995432 signifikante Paare ermittelt, in denen insgesamt 210162 verschiedene Wortformen vorkommen. Stoppwörter wurden aus den Betrachtungen herausgenommen. Sie sind als Listen sowie graphisch aufbereitet zugänglich unter [wortschatz.uni-leipzig.de](http://wortschatz.uni-leipzig.de) (SCHMIDT 1999, QUASTHOFF 1999a).

Die folgende Abbildung enthält die Kollokationen für *Schweine*:

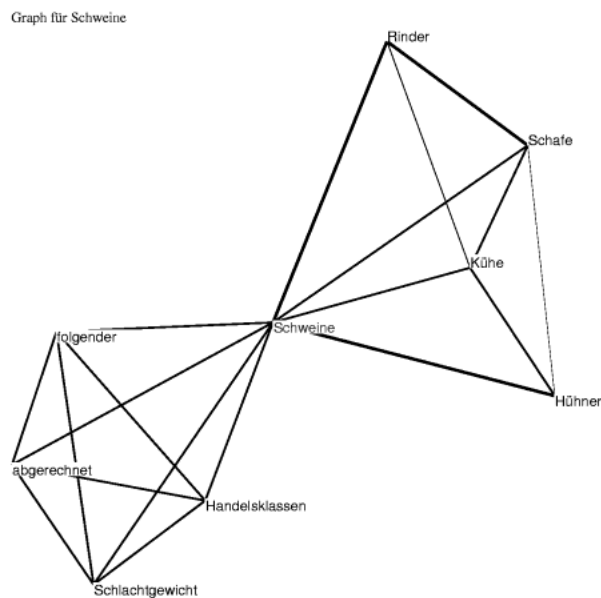


Abbildung 1: Kollokationen für *Schweine*

Die Abbildung zeigt, dass die Kollokationsmenge in zwei semantisch verschiedene Teile zerfällt: Rechts finden sich die Namen anderer Nutztierarten als Kohyponyme, links andere Wörter aus dem Umfeld. Diese Abbildung wurde zwar automatisch erstellt, aber mit solchen geometrischen Methoden lässt sich nicht erkennen, dass es sich rechts um Kohyponyme handelt, links dagegen nicht. Analog enthalten die Kollokationen zu *Katze* die Kohyponyme *Hund* und *Maus*, aber auch *Sack* und *gelassen*.

Um zunächst größere Gruppen unabhängig von ihrer semantischen Struktur identifizieren zu können, wurden zunächst Tripel von Kollokationen ermittelt, d. h. Tripel von Wortformen, die signifikant häufig gemeinsam im gleichen Satz auftreten. Danach wurde das Verfahren auch für Quadrupel bis Septupel angewendet. Dabei entstanden die folgenden Anzahlen von Tupeln:

<b>Tupelgröße</b>	<b>Anzahl</b>	<b>Beispiele</b>
Tripel	335.904	<i>Laufen Vorbereitungen Hochtouren rot Blüten gefärbten gelb rot grün sitzen essen trinken trinken Wein Gläser Licht Farbe Raum Farbe rot blau</i>
Quadrupel	388.036	<i>Meerschweinchen Katzen Hunde Kaninchen Schrank Stuhl Tisch Bett Bier Wein trinken Gläser</i>
Quintupel	703.677	<i>gelb orange rot grün blau Landgericht Rassenhaß Freiheitsstrafe Aufstachelung Bewährung</i>
Sextupel	1.263.541	<i>Lösemittel Laugen Farben Säuren Fette Sprays</i>
Septupel	2.038.283	<i>Landgericht Rassenhaß Freiheitsstrafe Aufstachelung Bewährung Volksverhetzung verurteilt; Stuttgart Elber Balakow Verlaat Foda Bobic Trautner</i>

Tabelle 1: Kollokationen im Satz

Es mag zunächst verblüffen, dass die Anzahlen bei wachsender Tupelgröße nicht drastisch abnehmen. Die Erklärung liegt in der Tatsache,

dass es in den verwendeten Texten gelegentlich zu längeren Aufzählungen solcher Kohyponyme kommt (z. B. Mannschaftsaufstellungen beim Fußball oder Mitteilungen zur Schadstoffsammlung). Aus einer solchen großen Gruppe mit beispielsweise 15 Elementen kann man 455 Tripel, aber 5005 Sextupel auswählen.

Die semantischen Beziehungen lassen sich in folgende Gruppen teilen:

- häufig gebrauchte Fügungen (z. B. *trinken Wein Gläser* aus *einige Gläser Wein trinken*);
- Kohyponymie (z. B. *gelb rot grün*; *Schrank Stuhl Tisch Bett*);
- mehrere Kohyponyme und ein Wort, welches mit allen assoziiert ist (z. B. *Farbe rot blau*; *Stuttgart Elber Balakow Verlaat Foda Bobic Trautner*);
- mehrere Wörter, die zwar in einem inhaltlichen Zusammenhang stehen, der sich aber nicht leicht automatisch erschließen lässt (z.B. *Landgericht Rassenhaß Freiheitsstrafe Aufstachelung Bewährung Volksverhetzung verurteilt*).

### Identifikation von Kohyponymen

Im folgenden soll ein Kriterium angegeben werden, welches es erlaubt, gute Kandidaten für die ersten beiden Gruppen zu finden. Schwierigkeiten machen offensichtlich Wörter, die mit einer größeren Anzahl von anderen Wörtern gemeinsam auftreten. Einerseits treten außer Kohyponymie hier häufig andere Relationen auf, andererseits ist dies wegen der großen Anzahl der entstehenden Tupel schwierig auszuwerten. Bemerkenswerterweise zeigen sich drastische Unterschiede in der Homogenität nicht an der Stärke der Signifikanz oder der Anzahl der Tripel, sondern der Veränderung der Tupelanzahl bei wachsender Größe. Tabelle 2 zeigt die Anzahl der Kollokationstupel, die jeweils ein bestimmtes Wort enthalten.

Die letzte Spalte enthält das Ergebnis, welches man aus den größten Tupeln und (in eckigen Klammern angegeben) durch eventuelle Erweiterung mit Elementen aus kleineren Tupeln (mit starker Überlappung) erhält. Beispielsweise erhält man das Ergebnis für *Aluminium* aus den zwei (sich überlappenden) Quintupeln, das Ergebnis zu *Tagebuch* enthält sechs Tripel, die sich nicht weiter überschneiden. Das Quintupel zu *gelb* erwei-

tert sich wie in eckigen Klammern angegeben, wenn die Quadrupel berücksichtigt werden.

Die in der Tabelle gezeigten Beispiele sowie weitere Tests legen die Vermutung nahe, dass ein schnelles Abklingen der Tupelzahlen bei wachsender Tupelgröße eine sinnvolle Einschränkung der näher zu untersuchenden Wörter darstellt. Mit dieser Regel schließen wir die letzten beiden Zeilen der Tabelle 2 von einer Untersuchung aus.

Wort	Paare	Tripel	Quadrupel	Quintupel	Sextupel	Septupel	Ergebnis
Aluminium	54	18	8	2	-	-	<i>Nickel, Kupfer, Zink, Blei, Aluminium, Zinn</i>
Tagebuch	52	6	-	-	-	-	<i>Tagebuch Che Guevara; Tagebuch Klemperer Victor; Tagebuch Dzevad Karahasan; Tagebuch Frank Anne; Tagebuch Ea Allesch; Tagebuch Jünger Ernst</i>
gelb	20	21	11	1	-	-	<i>gelb orange rot grün blau [braun Farben weiß schwarz rosa]</i>
orange	23	6	4	1	-	-	<i>gelb orange rot grün blau</i>
Ukraine	162	56	24	8	2	-	<i>Georgien Ukraine Aserbaid- schan Moldawien Armenien Rußland Bulgarien [Gruppe Vorrunde USA Türkei Rumänien]</i>
Aser- baid- schan	32	35	13	6	3	-	<i>Georgien Türkei Aserbaid- schan Moldawien Bulgarien Albanien Ukraine Armenien Rußland [Kasachstan Tadschikistan Usbekistan Kirgisien]</i>
Bewährung	100	236	156	44	9	2	(kein sinnvolles Ergebnis)
Freiheit	200	52	143	340	610	390	(kein sinnvolles Ergebnis)
Verlaa	47	262	1377	4715	24471	14358	(kein sinnvolles Ergebnis)

Tabelle 2: Verhalten der Tupelzahlen für Kohyponyme

Um die entstandenen Gruppen auf semantische Homogenität zu überprüfen und einzelne evtl. zusätzlich enthaltene Elemente (wie *Farben* bei *gelb* sowie *Gruppe* und *Vorrunde* bei *Ukraine*) zu eliminieren, können die dazu ermittelten Kollokations-*n*-Tupel ermittelt und mit dem jeweils vorhandenen Tupel verglichen werden. Große Unterschiede erlauben das sichere Aussondern des Elements.

### Identifizierung von Wendungen

Bisher wurde für das Signifikanzmaß das gemeinsame Auftreten im Satz benutzt. Analog lassen sich auch Wendungen identifizieren. Dies ist von vielfältigem Interesse, siehe z. B. SEELBACH (1999). Das Vorgehen ist folgendermaßen: Für Paare von Wörtern wird gezählt, wie oft sie unmittelbar nebeneinander auftreten. Überschreitet diese Häufigkeit eine Schwelle, werden rechte bzw. linke Nachbarn hinzugenommen, so dass jetzt Dreiergruppen benachbarter Wörter gezählt werden, anschließend Vierergruppen usw. Die Schwelle für die Häufigkeit wurde einheitlich mit 5 angesetzt. Anschließend wird aus der Häufigkeit einer Wortgruppe und den Einzelhäufigkeiten wie oben ein Signifikanzmaß ermittelt. Stoppwörter müssen hier berücksichtigt werden, da sie durchaus Teil der Wendung sein können (jedoch nicht als letztes Wort).

Tupelgröße	Anzahl	Beispiele
Tripel	2101963	<i>ARD und ZDF</i> <i>Jetzt oder nie</i> <i>Der Mann hatte</i>
Quadrupel	765750	<i>Auf die Palme gebracht</i> <i>Das Dumme ist nur</i> <i>Der Gewinn vor Steuern</i> <i>Ein Hund namens Beethoven</i>
Quintupel	198698	<i>Nicht mehr und nicht weniger</i> <i>Auf die schiefe Bahn geraten</i>
Sextupel	63860	<i>Der Internationale Gerichtshof in Den Haag</i> <i>Die drei Weisen aus dem Morgenland</i>
Septupel	28228	<i>Der Mann hatte nach Angaben der Polizei</i> <i>Gleich zwei Fliegen mit einer Klappe schlagen</i>

Tabelle 3: Kollokationen mit unmittelbaren Nachbarn

Insgesamt wurden so die in Tabelle 3 angegebenen Anzahlen von Wortgruppen extrahiert.

Die Beispiele in der Tabelle zeigen, dass folgende Arten von Wortgruppen auftreten:

- Feste Wendungen (in der Regel mit einer übertragenen Bedeutung), z. B. *auf die Palme gebracht*;
- Wiederholt auftretende Wortfolgen (ohne übertragene Bedeutung), die eine semantische Einheit bezeichnen und in dieser Form wiederholt auftreten (z. B. mehrgliedrige Eigennamen, bestimmte Aufzählungen und Präpositionalphrasen), z. B. *ARD und ZDF*;
- Andere wiederholt auftretende Wortfolgen, denen keine besondere Rolle zugesprochen werden soll, z. B. *Der Mann hatte*.

Auch hier stellt sich wieder das Problem, speziell Wortgruppen der letzten Kategorie auszusondern. Aus den vorliegenden Daten können wir einige Kriterien angeben, die bei einer Entscheidung helfen können, aber sicher nicht ausreichend sind.

- Typischerweise besteht eine Wendung aus wenigen Wörtern und ist maximal in dem Sinne, dass diese Wendung nicht Teil einer längeren ermittelten Wortgruppe ist. Dies trifft zu auf *ARD und ZDF*, aber nicht auf *der Mann hatte*.
- In vielen Fällen werden Wendungen bevorzugt am Anfang eines Satzes gebraucht. Allerdings muss die Wendung dies syntaktisch erlauben. (Dies ist der Fall für *jetzt oder nie*, aber nur bedingt für *auf die Palme gebracht*.)
- Mit einer syntaktischen Analyse lassen sich einige längere Wortgruppen sicher ausschließen, beispielsweise *Der Mann hatte nach Angaben der Polizei*.

Dass diese Regeln nicht ausreichend sind, zeigen beispielsweise Filmtitel, die alle drei Kriterien erfüllen können und trotzdem nicht als Wendungen akzeptiert werden sollten.

## Literatur

- DUNNING, T. (1994): Accurate Methods for the Statistics of Surprise and Coincidence. In: S. ARMSTRONG (ed.), *Using large Corpora*, 61-74. MIT Press.
- QUASTHOFF, U. (1999a): Der Deutsche Wortschatz im Internet, erscheint in: *LDV-Forum* 1999.
- QUASTHOFF, U. (1999b): Statistische Signifikanz für lexikalische Angaben. IfI-Report 1999, Universität Leipzig.
- SCHMIDT, F. (1999): Automatische Ermittlung semantischer Zusammenhänge lexikalischer Einheiten und deren graphische Darstellung. Diplomarbeit, Universität Leipzig.
- SEELBACH, D. (1999): Prädikative adjektivale Ausdrücke: Kodierung und kontrastive Aspekte Französisch-Deutsch (in diesem Band).
- WEBER, E. (1980): *Grundriss der biologischen Statistik*. Jena: Gustav Fischer Verlag.