Zusammenfassung

1. Motivation

Das Verb hat eine zentrale Position im Satz, da es die Struktur und die Bedeutung eines Satzes maßgeblich beeinflusst. Lexikalische Informationen zu Verben sind daher von grundlegender Bedeutung im Bereich der natürlichen Sprachverarbeitung. Es ist allerdings sehr aufwendig, die Feinheiten der natürlichen Sprache manuell zu definieren, und besonders semantische Ressourcen stellen in diesem Zusammenhang einen Engpass dar. Automatische Methoden zur Induktion von lexikalischem Wissen haben daher erheblich an Bedeutung gewonnen. Diese Arbeit stellt einen Beitrag zur automatischen Erstellung von lexikalisch-semantischem Wissen dar, der die Möglichkeiten und Grenzen einer Methode für eine automatische Induktion von semantischen Verbklassen fürs Deutsche vorstellt.

Semantische Verbklassen

Semantische Verbklassen generalisieren über Verben in Bezug auf ihre semantischen Eigenschaften. Die Klassen sind ein nützliches Mittel, umfangreiche Informationen zu Verben zu erfassen, ohne die idiosynkratischen Details für jedes Verb zu definieren. Die Namen der Verbklassen beschreiben die Verben auf einer konzeptuellen Ebene, und die idiosynkratischen lexikalischen Eigenschaften der Verben bleiben entweder unterspezifiziert oder werden hinzugefügt. Beispiele für Verbklassen auf der konzeptuellen semantischen Ebene sind *Positionsverben* (*liegen, sitzen, stehen*) oder *Bewegungsverben mit einem Fahrzeug* (*fahren, fliegen, rudern*). Für einige Sprachen sind bereits semantische Verbklassen definiert worden, z.B. Englisch (Levin, 1993; Baker *et al.*, 1998) und Spanisch (Vázquez *et al.*, 2000). Nach meinem Wissen existiert im Deutschen keine semantische Verbklassifikation, die maschinell für die natürliche Sprachverarbeitung zur Verfügung steht.

Was ist der Nutzen von Verbklassen in der natürlichen Sprachverarbeitung? Einerseits reduzieren Verbklassen Redundanz in Verbbeschreibungen, da sie sich auf gemeinsame Eigenschaften von Verben beschränken. Andererseits können Verbklassen Eigenschaften von Verben, die empirisch nur ungenügend beschrieben sind, hinzufügen oder vorhersagen, weil Eigenschaften von Verben derselben Klasse übertragen werden können. Es gibt bereits Arbeiten, die, basierend auf der englischen Verbklassifikation von Levin (1993), den Nutzen von Verbklassen in der natürlichen Sprachverarbeitung demonstriert haben: Dorr and Jones (1996) stellen einen Ansatz zur Desambiguierung von Verbbedeutungen vor, Dorr (1997) beschreibt den Nutzen der Klassifikation im

Bereich der maschinellen Übersetzung, und Klavans and Kan (1998) verwenden die Verbklassen für Dokumentenklassifikation.

Automatische Induktion von Semantischen Verbklassen im Deutschen

Wie können wir eine semantische Klassifikation von Verben erstellen, ohne sie von Hand zu definieren? Es ist schwierig, semantische Eigenschaften auf der Basis von vorhandenen Ressourcen zu lernen, sowohl in Bezug auf lexikalische Eigenschaften als auch in Bezug auf konzeptuelle Strukturen. Üblicherweise wird daher auf eine etablierte linguistische Hypothese zurückgegriffen, die einen engen Bezug zwischen den lexikalischen Bedeutungseigenschaften und dem Verhalten eines Verbs vorhersagt: Zu einem gewissen Grad bestimmt die Bedeutung eines Verbs sein Verhalten im Satz, besonders in Bezug auf die Wahl der Verbargumente, cf. Levin (1993, page 1). Diese Hypothese ist sehr nützlich bei der Induktion von semantischen Verbklassen, denn wir können eine Verbklassifikation aufgrund der Verhaltensweisen eines Verbs erstellen (die leichter zu lernen sind als semantische Eigenschaften), und diese Verhaltensklassifikation sollte zu einem gewissen Grad mit einer semantischen Klassifikation übereinstimmen.

Das Verbverhalten kann durch Verbalternationen dargestellt werden, alternative Verbkonstruktionen der Syntax-Semantik-Schnittstelle, die entweder dieselbe oder eine sehr ähnliche Bedeutung haben. Beispiel (1) stellt die häufigsten Alternationen für das Verb *fahren* dar. Die konzeptuellen Argumente, die typischerweise im Zusammenhang mit *fahren* verwendet werden, sind ein Fahrer, ein Fahrzeug, eine gefahrene Person und ein Weg. In (a) ist ein Fahrzeug das Subjekt der transitiven Verbalphrase, und eine Präpositionalphrase bezeichnet den Weg der Bewegung. In (b) ist ein Fahrer das Subjekt der transitiven Verbalphrase, und eine Präpositionalphrase bezeichnet wiederum den Weg der Bewegung. In (c) ist ein Fahrer das Subjekt der transitiven Verbalphrase, und das Akkusativobjekt bezeichnet ein Fahrzeug. In (d) ist ein Fahrer das Subjekt der ditransitiven Verbalphrase, das Akkusativobjekt bezeichnet eine gefahrene Person, und eine Präpositionalphrase bezeichnet den Weg.

(1) (a) *Der Wagen fährt in die Innenstadt.*

(b) *Die Frau fährt nach Hause.*

(c) *Der Filius fährt einen blauen Ferrari.*

(d) *Der Junge fährt seinen Vater zum Zug.*

Unter der Annahme, dass das Verbverhalten durch die Alternationen des Verbs charakterisiert werden kann, welche syntaktischen und semantischen Eigenschaften benötigt man für eine Beschreibung des Verbverhaltens? Beispiel (1) hat verdeutlicht, dass (i) die syntaktischen Strukturen relevant sind, weil sie die Funktionen der Verbargumente realisieren, (ii) Präpositionen relevant sind, um zum Beispiel Richtungsangaben von Ortsangaben zu unterscheiden, und (iii) Selektionspräferenzen relevant sind, weil sie die konzeptuellen Rollen der Arguments definieren. Diese drei Ebenen werden im Folgenden für eine Verbbeschreibung verwendet.

Wenn nun eine solche Beschreibung des Verbverhaltens vorliegt, wie kann man diese als Grundlage für die automatische Induktion von semantischen Verbklassen verwenden? Diese Arbeit

wendet einen Algorithmus zum Clustering an, der die Verbbeschreibung auf der Syntax-Semantik-Schnittstelle als empirische Verbeigenschaften benutzt und auf der Basis dieser Verbeigenschaften eine semantische Klassifikation lernt. Bei der Anwendung eines Clustering-Algorithmus auf multivariante Daten kommt es allerdings oft zu Überraschungen, weil man nicht notwendigerweise die Effekte der Datendefinition und des Algorithmus unterscheiden kann und das Ergebnis der Clusteranalyse daher nicht leicht zu interpretieren ist. Meiner Meinung nach ist es aber sehr wichtig, die Daten und den Algorithmus unter Berücksichtigung des Aufgabengebietes zu definieren. Wir forschen im Bereich der Linguistik, also sollten auch die Daten linguistisch definiert sein und der Algorithmus mit linguistischen Daten umgehen können. Im Rahmen dieser Arbeit habe ich mich daher auf zwei Teilaspekte der automatischen Induktion von Verbklassen konzentriert. Ich habe empirisch untersucht, wie Verben auf der Syntax-Semantik-Schnittstelle definiert werden können, mit anderen Worten: (i) Welches sind die semantischen Eigenschaften, die semantische Verbklassen charakterisieren? (ii) Welches sind die Eigenschaften, die Verbverhalten charakterisieren? (iii) In welchem Maße kann man den Zusammenhang zwischen Verbbedeutung und Verbverhalten benutzen, um semantische Verbklassen automatisch zu lernen? Der zweite Teilaspekt beschäftigt sich mit der Entwicklung einer Clustering-Methode, die in der natürlichen Sprachverarbeitung eingesetzt werden kann. Die Parameter des Algorithmus sollen so definiert werden, dass die Clusteranalyse weitgehend mit einer natürlichsprachlichen Klassifikation übereinstimmt. Meiner Meinung nach ist es sehr wichtig, das Potential und die Einschränkungen der linguistischen und technischen Teilaspekte zu verstehen; denn nur so kann eine Methode entwickelt werden, die uneingeschränkt auf Verbdaten angewendet werden kann.

2. Clustering-Methode

Die Clustering-Methode bringt die folgenden Aspekte in einen Zusammenhang: (a) das Konzept einer semantischen Verbklassifikation im Deutschen, (b) die empirischen Daten für die Verbbeschreibung auf der Syntax-Semantik-Schnittstelle, (c) den Clustering-Algorithmus und (d) Maße für die Evaluierung von Clusteranalysen.

Semantische Verbklassen im Deutschen

Ich habe manuell eine Verbklassifikation mit 43 semantischen Klassen und 168 teilweise ambigen deutschen Verben erstellt. Die Konstruktion der Verbklassen hat sich primär an den semantischen Eigenschaften der Verben orientiert: Die Verben werden den Klassen aufgrund ihrer lexikalischen und konzeptuellen Bedeutung zugewiesen, und die Klassen erhalten einen konzeptuellen Namen. Die Klassen beziehen sich auf zwei konzeptuelle Ebenen; allgemeine konzeptuelle Klassen wie z.B. *Bewegungsverben* wurden in spezifischere konzeptuelle Klassen unterteilt wie z.B. *Art der Fortbewegung, Rotation, Eile, Fahrzeug, Fließen*. Da es einen Bezug zwischen Verbbedeutung und Verbverhalten gibt, zeigen die Verben, die aufgrund der konzeptuellen Bedeutung einer Klasse zugewiesen wurden, auch eine gewisse Übereinstimmung in ihrem Verhalten. Die Klassengröße in der Verbklassifikation variiert von 2 bis 7 Verben pro Klasse, was einem Durchschnitt von 3.9 Verben pro Klasse entspricht. Acht Verben sind in Bezug auf die Klassenzuweisung ambig und wurden jeweils zwei Klassen zugeordnet. Die Klassen enthal-

ten sowohl hochfrequente als auch niedrigfrequente Verben: Die Verbfrequenz variiert von 8 bis 71.604. Die Verbklassifikation entspricht bei der Evaluierung der Clustering-Experimente dem goldenen Standard.

Jede Klasse wurde detailliert im Rahmen von Fillmore's 'Scenes-and-Frames' Semantik (Fillmore, 1977, 1982) beschrieben, die in ähnlicher Weise bereits im Rahmen von *FrameNet* verwendet wurde (Baker *et al.*, 1998; Johnson *et al.*, 2002): Eine Szenenbeschreibung charakterisiert die Klasse und definiert die Rollen für obligatorische und modifizierende Verbargumente. Die Verbrahmen, die im Rahmen der Klassenbedeutung Verwendung finden, werden aufgelistet. Die Rahmen und die Rollen wurden auf der Basis von Zeitungskorpora aus den 1990er Jahren entwickelt. Für jedes Verb und jeden Rahmen in einer Klasse wurden Beispielsätze aus den Korpora extrahiert. Da in jedem Satz Rahmen und Rollen annotiert wurden, illustrieren die Beispielsätze die Verbalternationen, denn die Rollen können über die Rahmengrenzen hinweg in Verbindung gebracht werden.

Empirische Distributionen für deutsche Verben

Ich habe ein statistisches Grammatikmodell fürs Deutsche entwickelt, implementiert und trainiert. Die Grammatik beruht auf der Idee von lexikalisierten probabilistischen kontextfreien Grammatiken, die ursprünglich von Charniak (1995) entwickelt wurde. Im Rahmen dieser Arbeit habe ich eine Implementierung von Schmid (2000) verwendet. Das statistische Grammatikmodell enthält empirische lexikalische Information, die auf die Subkategorisierung von Verben spezialisiert, aber nicht beschränkt ist.

Die empirische Information in dem Grammatikmodell wird für die Verbbeschreibung verwendet. Ich stelle die deutschen Verben durch distributionelle Vektoren dar, basierend auf der Hypothese von Harris (1968), dass 'jede Sprache durch ihre distributionelle Struktur beschrieben werden kann, d.h. durch das Vorkommen mit anderen Teilen der Sprache'. Die Verben werden distributionell auf drei Ebenen der Syntax-Semantik-Schnittstelle beschrieben, wobei jede Ebene die vorherige durch zusätzliche Information anreichert. Auf der Ebene $D1$ werden rein syntaktische Eigenschaften von Verben dargestellt, auf der Ebene $D2$ wird präpositionale Information hinzugefügt, und auf der Ebene $D3$ werden Selektionspräferenzen definiert. Ich fange so auf einer rein syntaktischen Ebene für Verbbeschreibung an und füge schrittweise semantische Information hinzu. Die umfangreichste Verbbeschreibung entspricht in etwa der Beschreibung von Verbalternation. Ich habe mich für diese Dreiteilung von Verbeigenschaften entschieden, weil die Clusteranalysen auf der Basis der verschiedenen Verbeigenschaften Einblick in den Zusammenhang zwischen Verbbedeutung und Verbverhalten geben sollen.

Die folgende Tabelle illustriert die Verbbeschreibungen auf den drei Ebenen, und zwar für drei Verben aus drei unterschiedlichen Verbklassen. Für jede Ebene der Verbbeschreibung werden die zehn wahrscheinlichsten Verbeigenschaften dargestellt. Die Rahmen der Verben setzen sich aus Abkürzungen für Verbargumente zusammen: Nominalphrasen im Nominativ (n), Akkusativ (a) und Dativ (d), Reflexivpronomen (r), Präpositionalphrasen (p), Expletive (x), nicht-finite subkategorisierte Sätze (i), finite subkategorisierte Sätze (s-2 bei Verb-Zweit-Sätzen, s-dass bei *dass*-Sätzen, s-ob bei *ob*-Sätzen und s-w bei indirekten Fragen) und Kopula-Konstruktionen (k). Der

konkrete Bezug auf Präpositionalphrasen erfolgt durch Kasus und Präposition, z.B. 'mit$_{Dat}$', und 'für$_{Akk}$'. Die Selektionspräferenzen werden durch die 15 Top-Knoten in *GermaNet* (Hamp and Feldweg, 1997; Kunze, 2000) dargestellt: *Lebewesen, Sache, Besitz, Substanz, Nahrung, Mittel, Situation, Zustand, Struktur, Physis, Zeit, Ort, Attribut, Kognitives Objekt, Kognitiver Prozess*. Ich habe die Selektionspräferenzen dadurch definiert, dass ich Frequenzen von nominalen Fillern in bestimmten Argumentpositionen aufwärts durch die GermaNet-Hierarchie propagiert und dann eine Verallgemeinerung auf der höchsten Ebene abgelesen habe. Das Argument in einem Rahmen, das durch Selektionspräferenzen ergänzt wird, ist jeweils unterstrichen. Die Kerninformation bei den Verbbeschreibungen, die Rahmeninformation auf den Ebenen $D1$ und $D2$, ist evaluiert worden. In Schulte im Walde (2002b) habe ich ein Subkategorisierungslexikon für 14.229 Verben aus dem statistischen Grammatikmodell abgeleitet, mit Verbfrequenzen von 1 bis 255.676. Eine Evaluierung des Lexikons in Schulte im Walde (2002a) hat gezeigt, dass die Subkategorisierungsinformation manuelle Definitionen ergänzen und verbessern kann und daher wertvoll für einen Einsatz in der natürlichen Sprachverarbeitung ist.

Verb	Distribution					
	$D1$		$D2$		$D3$	
beginnen	np	0.43	n	0.28	n(Situation)	0.12
	n	0.28	np:um$_{Akk}$	0.16	np:um$_{Akk}$(Situation)	0.09
	ni	0.09	ni	0.09	np:mit$_{Dat}$(Situation)	0.04
	na	0.07	np:mit$_{Dat}$	0.08	ni(Lebewesen)	0.03
	nd	0.04	na	0.07	n(Zustand)	0.03
	nap	0.03	np:an$_{Dat}$	0.06	np:an$_{Dat}$(Situation)	0.03
	nad	0.03	np:in$_{Dat}$	0.06	np:in$_{Dat}$(Situation)	0.03
	nir	0.01	nd	0.04	n(Zeit)	0.03
	ns-2	0.01	nad	0.03	n(Sache)	0.02
	xp	0.01	np:nach$_{Dat}$	0.01	na(Situation)	0.02
essen	na	0.42	na	0.42	na(Lebewesen)	0.33
	n	0.26	n	0.26	na(Nahrung)	0.17
	nad	0.10	nad	0.10	na(Sache)	0.09
	np	0.06	nd	0.05	n(Lebewesen)	0.08
	nd	0.05	ns-2	0.02	na(Lebewesen)	0.07
	nap	0.04	np:auf$_{Dat}$	0.02	n(Nahrung)	0.06
	ns-2	0.02	ns-w	0.01	n(Sache)	0.04
	ns-w	0.01	ni	0.01	nd(Lebewesen)	0.04
	ni	0.01	np:mit$_{Dat}$	0.01	nd(Nahrung)	0.02
	nas-2	0.01	np:in$_{Dat}$	0.01	na(Attribut)	0.02
fahren	n	0.34	n	0.34	n(Sache)	0.12
	np	0.29	na	0.19	n(Lebewesen)	0.10
	na	0.19	np:in$_{Akk}$	0.05	na(Lebewesen)	0.08
	nap	0.06	nad	0.04	na(Sache)	0.06
	nad	0.04	np:zu$_{Dat}$	0.04	n(Ort)	0.06
	nd	0.04	nd	0.04	na(Sache)	0.05
	ni	0.01	np:nach$_{Dat}$	0.04	np:in$_{Akk}$(Sache)	0.02
	ns-2	0.01	np:mit$_{Dat}$	0.03	np:zu$_{Dat}$(Sache)	0.02
	ndp	0.01	np:in$_{Dat}$	0.03	np:in$_{Akk}$(Lebewesen)	0.02
	ns-w	0.01	np:auf$_{Dat}$	0.02	np:nach$_{Dat}$(Sache)	0.02

Auf $D1$ sind die wahrscheinlichsten Rahmen für *beginnen* ‘np’ und ‘n’. $D2$ zeigt, dass auch nach dem Verteilen der Rahmenwahrscheinlichkeit für ‘np’ über die verschiedenen Arten von PPs noch eine Reihe von prominenten PPs zu finden ist, temporal um_{Akk}, an_{Dat} und $nach_{Dat}$, die Kennzeichnung eines beginnenden Ereignisses durch mit_{Dat}, lokative PPs mit in_{Dat}. $D2$ macht deutlich, dass nicht nur PPs in Argumentfunktion, sondern auch PPs in Adjunktfunktion eine wichtige Rolle im Verbverhalten darstellen. $D3$ zeigt, dass typische beginnende Ereignisse durch *Situation, Zustand, Zeit, Sache* charakterisiert werden. Außerdem kann man die implizite Definition von Alternationsverhalten erkennen, denn ‘na(Situation)’ in einem transitiven Verbrahmen bezieht sich auf die gleiche Rolle wie ‘n(Situation)’ in einem intransitiven Verbrahmen. Das Verb *essen* zeigt starke Präferenzen für sowohl einen transitiven als auch einen intransitiven Rahmen. Wie gewünscht bestimmen *Lebewesen* in ‘n’ und ‘na’ sowie *Nahrung* in ‘na’ die Selektionspräferenzen und das Alternationsverhalten. *fahren* taucht mit den typischen Rahmen für Bewegungsverben auf: ‘n’, ‘np’, ‘na’. Die PPs beziehen sich entweder auf eine Richtungsangabe (in_{Akk}, zu_{Dat}, $nach_{Dat}$) oder auf ein Fahrzeug (mit_{Dat}, in_{Dat}, auf_{Dat}). Auch hier weisen die Selektionspräferenzen auf das typische Alternationsverhalten hin: *Lebewesen* in ‘n’ und ‘na’, *Sache* in der kausativen Alternation ‘n’ und ‘na’.

Algorithmen für eine Clusteranalyse und ihre Evaluierung

Die Clusteranalyse für die deutschen Verben wurde durch den k-Means Algorithmus durchgeführt (Forgy, 1965). k-Means ist ein Standard-Algorithmus im Bereich des Clustering, der iterativ Cluster re-organisiert und Verben ihrem jeweils nächsten Cluster-Mittelpunkt zuweist, bis es keine Änderungen mehr gibt. Eine Anwendung des k-Means Algorithmus kommt den Annahmen gleich, (i) dass die Verben durch distributionelle Vektoren beschrieben werden und (ii) dass Verben, die aufgrund mathematischer Berechnung nicht weit voneinander entfernt sind, auch im linguistischen Sinn einander ähnlich sind.

k-Means erfordert die Definition von einer Reihe von Parametern: Die Anzahl der Cluster in der Clusteranalyse ist nicht gegeben und muss daher experimentell ermittelt werden. In engem Bezug zu diesem Parameter steht die Ebene der konzeptuellen Struktur: Je mehr Cluster in der Clusteranalyse vorhanden sind, desto spezifischer ist die konzeptuelle Ebene. Außerdem kann die Eingabe für den Algorithmus variiert werden: Es können Zufallscluster vorgegeben werden, oder die Eingabe kann durch eine vorhergehende Clusteranalyse vorverarbeitet werden. Die Clustereingabe ist von entscheidender Bedeutung bei k-Means, und daher habe ich sowohl Experimente mit Zufallsclustern als auch mit vorverarbeiteten hierarchischen Clustern durchgeführt. Bei hierarchischen Clustern habe ich außerdem mit verschiedenen Arten der Cluster-Zusammenführung experimentiert. Ein weiterer Parameter bei k-Means ist das Distanzmaß zwischen Verben und Clustern. Welches Maß ist optimal, um Distanzen zwischen Verben zu berechnen? Die Clustering-Methode wurde auf der Basis der Verben in der manuellen Klassifikation erarbeitet und anschließend im Hinblick auf eine große Datenmenge von Verben getestet und diskutiert.

Um die Clusteranalysen zu evaluieren, braucht man ein unabhängiges und zuverlässiges Maß für die Bewertung und den Vergleich von Experimenten und Cluster-Ergebnissen. Theoretisch hat

die experimentierende Person ein Gefühl für die Güte der Ergebnisse, aber praktisch gesehen sind die Datenmengen zu groß und die Details in der Datenbeschreibung zu fein für eine objektive Bewertung. Es gibt keine Standard-Maße für die Evaluierung von Clusteranalysen, aber es gibt ähnliche Problemstellungen in anderen Forschungsbereichen wie z.B. der theoretischen Statistik, Bildverarbeitung und Clustering von Webseiten, deren Evaluierungsmethoden sich auf Verbklassen übertragen lassen. Ich habe eine Anzahl von generellen Anforderungen an eine Evaluierung, generellen Anforderungen an eine Clusteranalyse und spezifischen Anforderungen an eine linguistische Clusteranalyse formuliert und verschiedene Evaluierungsmaße im Hinblick auf diese Anforderungen bewertet. Als Ergebnis habe ich drei Evaluierungsmaße für die Experimente ausgewählt: Eine Berechnung von Precision- und Recall-Werten, die auf einer paarweisen Evaluierung von Verben beruht (Hatzivassiloglou and McKeown, 1993) und intuitiv einfach interpretiert werden kann, eine Berechnung von Precision-Werten, die ebenfalls auf einer paarweisen Evaluierung von Verben beruht, aber im Hinblick auf die linguistische Aufgabenstellung durch einen Skalierungsfaktor der Clustergröße optimiert wurde (Schulte im Walde and Brew, 2002) und die Berechnung des angepassten Rand-Index, der die Übereinstimmung und die Unterschiede von Cluster-Zuweisungen bestimmt und einen direkten Bezug zum Nullmodell beinhaltet (Hubert and Arabie, 1985).

Beispiele zu Clusteranalysen

Zu Illustrationszwecken stelle ich Beispiele von Clusteranalysen vor. Die erste Analyse klassifiziert die 168 deutschen Verben der manuell erstellten Klassifikation und beruht auf folgenden Parametern: Die Eingabe ist eine hierarchische Clusteranalyse, bei der die Cluster auf der Basis von *Ward's* Methode zusammengeführt wurden. Die Anzahl der Cluster ist die Anzahl der manuell erstellten Klassen (43), und Distanzmaße wurden durch die Skew-Divergenz, eine Variante der Kullback-Leibler-Divergenz, durchgeführt. Im Folgenden beschreibe ich repräsentative Teile der Clusteranalyse, die auf Verbbeschreibungen der Ebene $D3$ beruht, mit Selektionspräferenzen für die Rollen in 'n', 'na', 'nd', 'nad', 'ns-dass'. Ich vergleiche die ausgewählten Cluster mit den entsprechenden Clustern auf den Ebenen $D1$ und $D2$. In jedem Cluster werden alle Verben, die zu einer gemeinsamen konzeptuellen Verbklasse gehören, in einer Zeile aufgelistet zusammen mit den Klassennamen.

(a) beginnen enden – *Aspekt*
bestehen existieren – *Existenz*
liegen sitzen stehen – *Position*
laufen – *Bewegung: Art*

(b) kriechen rennen – *Bewegung: Art*
eilen – *Bewegung: Eile*
gleiten – *Bewegung: Fließen*
starren – *Gesichtsausdruck*

(c) klettern wandern – *Bewegung: Art*
fahren fliegen segeln – *Bewegung: Fahrzeug*
fließen – *Bewegung: Fließen*

(d) festlegen – *Festlegung*
bilden – *Produktion*
erhöhen senken steigern vergrößern verkleinern – *Maßänderung*

(e) töten – *Eliminierung*
unterrichten – *Lehre*

(f) nieseln regnen schneien – *Wetter*

(g) dämmern – *Wetter*

Die Wetterverben in Cluster (f) zeigen eine starke Übereinstimmung in ihren syntaktischen Rahmen auf $D1$. Sie werden daher schon auf der Basis von $D1$ einem gemeinsamen Cluster zugewiesen und brauchen keine Verfeinerungen auf $D2$ und $D3$. *dämmern* in Cluster (g) ist ein ambiges Verb. Es ist ein Wetterverb und sollte daher mit den Verben in Cluster (f) vorkommen, aber es hat auch eine Bedeutung von Verstehen (*mir dämmert ...*), die sich in einem Rahmen wiederfindet, der *dämmern* von den anderen Wetterverben unterscheidet. Es ist daher weder auf $D1 - D3$ mit den anderen Wetterverben in einem Cluster. Die Verben aus den Klassen *Bewegung, Existenz, Position* und *Aspekt* sind sich auf der Ebene $D1$ sehr ähnlich und können dort nicht unterschieden werden. Auf $D2$ unterscheiden die klassenspezifischen Präpositionen die Verben und erstellen eigene Cluster für die jeweiligen Klassen: Bewegungsverben verwenden typischerweise direktionale PPs, Aspektverben geben das beginnende Ereignis durch mit_{Dat} an oder definieren temporale oder lokative PPs, und Existenz- und Positionsverben verwenden lokative PPs, wobei die Positionsverben eine größere Variation zeigen. Die PP-Information ist relevant für die Unterscheidung der Verben, und auf $D2$ werden die Verben erfolgreich unterschieden. Aber diese Koheränz wird teilweise durch $D3$ zerstört: Die Mehrheit der Bewegungsverben (von den verschiedenen Unterklassen) ist in den Clustern (b) und (c), aufgrund einer starken Ähnlichkeit im Alternationsverhalten. Aber die Existenz-, Positions- und Aspektverben haben idiosynkratische Anforderungen an Selektionspräferenzen, so dass ihre Ähnlichkeit auf $D2$ durch die Verfeinerung auf $D3$ zerstört wird. Sie befinden sich alle in Cluster (a), dem man tatsächlich immer noch eine gewisse Ähnlichkeit nicht verwehren kann, denn alle Verben beschreiben eine Art von Existenz. In Cluster (d) sind fast alle Verben der *Maßänderung* zusammen mit einem Produktionsverb und einem Verb der *Festlegung*. Das Cluster ist also konzeptionell sehr gut. Die Verben in dem Cluster subkategorisieren typischerweise ein direktes Objekt oder 'nr' und 'npr', im letzteren Fall mit den PPs auf_{Akk} und um_{Akk}. Die Selektionspräferenzen sind entscheidend für die Erstellung dieses Clusters: Die Verben stimmen überein in einer Sache oder Situation als Subjekt und verschiedenen Objekten wie z.B. Attributen, kognitiven Objekten, Zuständen, Strukturen oder Dingen. $D1$ und $D2$ können diese Verben nicht unterscheiden. Schließlich gibt es Cluster wie in (e), deren Verben nur auf einer ganz allgemeinen konzeptuellen Ebene übereinstimmen. Bei *töten* und *unterrichten* erfolgt beispielsweise eine Aktion, die von einer Person oder Gruppe auf eine andere Person oder Gruppe gerichtet ist.

Eine zweite Clusteranalyse hat sich derselben Parameter bedient, aber eine größere Anzahl von Verben klassifiziert: Ich habe alle deutschen Verben mit einer empirischen Frequenz zwischen 500 und 10.000 im Trainingskorpus extrahiert. Die Gesamtheit von 809 Verben schließt 94

Verben der manuellen Klassifikation mit ein. Ich habe die fehlenden Verben der manuellen Klassifikation aus Evaluierungsgründen hinzugefügt, so dass sich eine Gesamtzahl von 883 Verben ergeben hat. Die Verben wurden in 100 Cluster eingeteilt, was einer durchschnittlichen Anzahl von 8.83 Verben pro Cluster entspricht. In der Clusteranalyse sind einige semantisch sehr gute Cluster, einige Cluster enthalten semantisch ähnliche und entferntere Verben, und bei einigen Clustern ist es sehr schwierig, überhaupt eine konzeptuelle Ähnlichkeit festzustellen. Für jede Art von Clustern folgen hier einige Beispiele. Verben, die ich innerhalb eines Clusters für sehr ähnlich halte, sind durch Fettdruck markiert.

(a) *anhören, auswirken, einigen, lohnen, verhalten, wandeln*

(b) *beschleunigen,* ***bilden****, darstellen, decken, erfüllen,* ***erhöhen****, erledigen, finanzieren, füllen, lösen, rechtfertigen,* ***reduzieren****,* ***senken****,* ***steigern****,* ***verbessern****,* ***vergrößern****,* ***verkleinern****,* ***verringern****,* ***verschieben****,* ***verschärfen****,* ***verstärken****,* ***verändern***

(c) ***ahnen****,* ***bedauern****,* ***befürchten****,* ***bezweifeln****,* ***merken****,* ***vermuten****, weißen,* ***wissen***

(d) ***anbieten****, angebieten,* ***bieten****,* ***erlauben****,* ***erleichtern****,* ***ermöglichen****,* ***eröffnen****,* ***untersagen****, veranstalten,* ***verbieten***

(e) ***basieren****,* ***beruhen****,* ***resultieren****,* ***stammen***

(f) ***befragen****,* ***entlassen****,* ***ermorden****,* ***erschießen****,* ***festnehmen****,* ***töten****,* ***verhaften***

Cluster (a) ist ein Beispiel, in dem die Verben keinerlei semantische Ähnlichkeit aufweisen. In der Clusteranalyse sind solche semantisch inkohärenten Cluster typischerweise sehr groß, mit 15-20 Verben pro Cluster. Cluster (b) ist ein Beispiel, in dem ein Teil der Verben eine semantische Ähnlichkeit aufweist, aber auch entferntere Verben enthalten sind. Die Mehrzahl der Verben gehören zur Klasse *Maßänderung*. Cluster (c) bis (f) sind Beispiele für semantisch sehr kohärente Klassen. Cluster (c) enthält Verben, die eine propositionale Einstellung formulieren; die unterstrichenen Verben beschreiben zudem noch eine Emotion. Auch das unmarkierte Verb *weißen* hat eine Berechtigung in dem Cluster, denn es beruht auf einem Fehler in der Morphologie, die Wortformen mit den Lemmata *wissen* und *weißen* nicht immer unterscheiden kann. Cluster (d) beschreibt eine Situation, in der jemand jemandem etwas ermöglicht sowohl im negativen als auch im positiven Sinn. Neben dem auf einem Lemmatisierungsfehler beruhenden *angebieten* (kein Infinitiv, sondern das inkorrekte Partizip Perfekt von *anbieten*) ist der einzige Fehler in dem Cluster das Verb *veranstalten*. In Cluster (e) beziehen sich alle Verben auf eine Basis, und in Cluster (f) beschreibt die Gesamtheit der Verben den Prozess von der Verhaftung eines Verdächtigen bis zu seiner Bestrafung. Die Clusteranalyse könnte nach einer manuellen Korrektur als lexikalische semantische Ressource verwendet werden.

3. Folgerungen

Ich habe eine Methodik für eine Clusteranalyse von deutschen Verben vorgeschlagen, deren Ergebnis mit einer manuellen Klassifikation in großen Teilen übereinstimmt und daher als automatische Basis für eine semantische Klassifikation von Verben verwendet werden kann. Die

Kernaspekte der Methodik beziehen sich einerseits auf die linguistischen und andererseits auf die technischen Aspekte der Aufgabenstellung.

Linguistische Aspekte

Ich habe eine Strategie verwendet, die Verbverhalten durch syntaktische Subkategorisierungseigenschaften, Präpositionen und Selektionspräferenzen beschreibt. Diese Strategie hat sich als sehr erfolgreich erwiesen, da sie den engen Zusammenhang zwischen Verbverhalten und Verbbedeutung darstellt, indem sie auf der Basis der Verbbeschreibungen semantische Klassen erzeugt. Die Clusteranalysen wurden durch jede Ebene der Verbbeschreibung verbessert. Die Auswertungen der Experimente haben die praktischen Grenzen der linguistischen Eigenschaften aufgezeigt: Es gibt eine linguistische Grenze, die durch die Unterscheidung von gemeinsamen und idiosynkratischen Verbeigenschaften definiert wird. Aus theoretischer Sicht ist einleuchtend, dass die Verben in einer gemeinsamen Klasse nur aufgrund ihrer gemeinsamen Eigenschaften in ein Cluster definiert werden können. Wenn aber die Verbbeschreibung schrittweise verfeinert wird und dadurch mehr und mehr idiosynkratische Eigenschaften eingeschlossen werden, schwindet die Gemeinsamkeit der Verben. Aus praktischer Sicht ist es sehr schwierig, die Grenze zwischen gemeinsamen und idiosynkratischen Eigenschaften festzulegen, weil sie von der konzeptuellen Ebene einer semantischen Klasse abhängt und daher von Klasse zu Klasse auf unterschiedlicher Ebene liegen kann. Die herausgearbeitete Definition von Subkategorisierung, Präpositionen und einer Auswahl von Selektionspräferenzen in dieser Arbeit hat sich dabei als ein linguistisch brauchbarer Kompromiss herausgestellt.

Technische Aspekte

Ich habe den Zusammenhang zwischen der grundlegenden Idee eines Clustering-Algorithmus, dessen Parametern und den resultierenden Clusteranalysen im Hinblick auf eine natürlichsprachliche Klassifikation untersucht. Die Eingabe für den Algorithmus spielt eine große Rolle, denn k-Means benötigt für eine linguistisch aussagekräftige Analyse kompakte Cluster mit ähnlicher Größe. Die passenden Analysen durch k-Means werden auf der Basis von denjenigen vorverarbeiteten hierarchischen Clustern erstellt, die wie k-Means eine Minimierung von der Distanz zwischen Verben und Cluster-Mittelpunkten durchführen. *Ward's* Methode hat sich dabei als die beste erwiesen und ist tatsächlich ähnlich erfolgreich auch ohne Nachverarbeitung durch k-Means. Die Ähnlichkeitsmaße von Verben auf der Basis ihrer Darstellung als Vektoren haben nur sekundäre Bedeutung. Erst bei großen Mengen an Objekten zum Clustern oder Eigenschaften für die Objektbeschreibung sind Varianten der Kullback-Leibler-Distanz bevorzugt. Für die distributionelle Beschreibung der Verben eignen sich sowohl Frequenzen als auch Wahrscheinlichkeiten; beide Arten der Darstellung können durch Smoothing noch verbessert werden. Die Anzahl der Cluster in der Clusteranalyse ist nur relevant in Bezug auf die Größenordnung. Eine große Anzahl von Clustern zu verlangen ist sehr ergeizig. Bevorzugt sollten die Verben weniger Clustern mit allgemeinerem Inhalt zugeordnet werden, da das die Fehlerrate senkt und die Cluster eine gute Ebene für manuelle Korrektur darstellen.