Author built version of : Diesner, J., & Carley, K. M. (2010) Relation Extraction from Texts (in German, title: Extraktion relationaler Daten aus Texten). In C. Stegbauer & R. Häußling (Eds.), Handbook Network Research (Handbuch Netzwerkforschung). Vs Verlag.

# Extraktion relationaler Daten aus Texten

*Jana Diesner, Kathleen M. Carley*

## 1 Einleitung

Daten für netzwerkanalytische Projekte können explizit oder implizit in natürlichsprachlichen, un- oder halbstrukturierten Texten enthalten sein. In dieser Situation ermöglichen Verfahren zur Relationsextraktion die Gewinnung oder Anreicherung von Netzwerkdaten. Die folgenden Beispiele verdeutlichen Einsatzgebiete für diese Familie von Methoden: Analysten aus Wirtschaft und Verwaltung entnehmen Berichten von und über Organisationen Angaben zu deren Zusammensetzung, Effizienz und Entwicklung (Corman et al. 2002; Krackhardt 1987). Kognitions- und Sozialwissenschaftler untersuchen auf der Grundlage von Interviews, wer welche Themen anspricht und wie in Verbindung setzt (Carley und Palmquist 1991; Collins und Loftus 1975). Journalisten und Analysten durchsuchen Meldungen und Archive nach Beteiligten, Gegenstand, Grund, Verlauf, Ort, Zeit, und Zusammenhängen von Ereignissen (Gerner et al. 1994; van Cuilenburg et al. 1986). Marktforscher analysieren Kundenbewertungen um herauszufinden, welche Marken und Produkte welche Empfindungen hinterlassen (Wiebe 2000). Internetforscher verfolgen die akteursbezogene Diffusion von Themen im Internet (Adar und Adamic 2005; Kleinberg 2003). Nutzer senden Suchmaschinen Anfragen, deren Beantwortung Informationen von mehr als einer Webseite bedarf (Berners-Lee et al. 2001; Brin 1999). All diesen Aufgaben ist gemeinsam, dass sie gelöst werden können, indem die jeweils relevanten Informationen (Knoten) und deren Verbindungen (Kanten) aus Texten herausgefunden, wiedergegeben und netzwerkanalytisch ausgewertet werden (McCallum 2005). In diesem Kapitel erläutern wir, unter welchen Bedingungen das Extrahieren relationaler Daten aus Texten sinnvoll ist, welche Verfahren dafür zur Verfügung stehen, und zeigen Grenzen und bislang ungelöste Probleme der Methodik auf.

## 2 Möglichkeiten und Grenzen des Einsatzes von Relationsextraktion

Die Extraktion relationaler Daten aus Texten ermöglicht zum einen das Erfassen von klassischen sozialen Netzwerkstrukturen. Zum Beispiel kann bei der Analyse von Emails aus den Emailköpfen entnommen werden, wer mit wem kommuniziert, was im Ergebnis ein traditionelles soziales Netz ergibt. Darüber hinaus erlauben Relationsextraktionsverfahren, Netzwerkdaten um das anzureichern, was durch ein Netz hindurchfließt. Diese zweite Dimension von Netzwerkdaten ist dann wertvoll, wenn die im Netz transportierte Materie zusätzliche Erkenntnisse über das zu untersuchende System ermöglicht. „Travelling through the network are fleets of social objects“ (Danowski 1993: 198). Diese Materie oder sozialen Objekte können unter anderem Güter, Viren, Informationen, oder Emotionen sein. Beim Beispiel der Emailnetzwerke kann das soziale Netz mit einem semantischen Netz, das aus den Emailbodys extrahiert wird, fusioniert und erweitert werden. Dabei entsteht ein

multi-modales Netz, mit dem untersucht werden kann, wer was mit wem kommuniziert. Wir verwenden den Begriff des semantischen Netzes hier recht allgemein als Bezeichnung für das Ergebnis der Verlinkung von relevanten Informationen aus Texten in eine Struktur. Diesner et al. (2007) haben für die interne Kommunikation von Enron, einem Konzern mit zweifelhaften Geschäftsmethoden, gezeigt, dass die Struktur und Merkmale sozialer und semantischer Netze, die aus gleicher Quelle (Emails) erhoben wurden, stark voneinander abweichen und somit Einsichten in verschiedene Aspekte sozialer Dynamik vermitteln können. Corman et al. (2002: 164) fassen diesen Punkt treffend zusammen: „We cannot reduce communication to message transmission".

Eine weitere Stärke relationaler Textanalysen ist die Berücksichtigung des Kontextes von Aussagen (Carley 1997; Collins und Loftus 1975; Janas und Schwind 1979). Die Methodik bietet damit eine Ergänzung und Erweiterung zu Verfahren, die den Inhalt von Texten erfassen, indem sie Worte oder deren Zuordnung zu Kategorien als von anderen Worten oder Kategorien konditional unabhängige Datenpunkte betrachten und deren jeweilige kumulative Häufigkeit ermitteln und vergleichen, wie z.B. bei der Inhaltsanalyse (Berelson 1952; Krippendorff 2004; Van Atteveldt 2008). Die dabei resultierenden Daten bilden ein Netz ohne Kanten, auf das jedes Netz im Bedarfsfall reduziert werden kann. Bedeutungsunterschiede, die nicht in der Identität und Häufigkeit von Knoten begründet liegen, sondern in deren Verlinkung und resultierenden Position im Netzwerk, können dabei nicht aufgedeckt werden. Die Kontextualisierung von Informationen, die sich durch die Berücksichtigung von Relationen ergibt, ermöglicht es, den Beziehungen zwischen Worten (Syntax), dem sozialen Gebrauch von Sprache (Pragmatik) und der Bedeutung von Aussagen (Semantik) näher zu kommen (Bernard und Ryan 1998; Carley und Palmquist 1992; Doerfel 1998; Mohr 1998; Woods 1975). Damit kann man Texte, deren Struktur und deren Bedeutung mit mikroskopischem (einzelne Worte, Knoten, Kanten) und makroskopischem (Triaden, Cluster, Netze) Blick untersuchen, zwischen diesen Perspektiven wechseln, Symbole bzw. Daten in Informationen und Wissen überführen, und Fehlinterpretationen komplexer Zusammenhänge reduzieren.

Relationsextraktion aus Texten ist dann eine sinnvolle Ergänzung oder Alternative zu klassischen Verfahren der Netzwerkdatengewinnung, wenn Informationen über ein zu untersuchendes System nicht mit traditionellen Verfahren wie Fragebögen, Beobachtungen oder Datenbankabfragen erhoben werden können, aber Textdaten zu dem System vorliegen. Das kann der Fall sein bei:

- Verdeckt agierenden und illegalen Netzen wie Wirtschaftskriminalität (Baker und Faulkner 1993) und Terrorgruppen (Krebs 2002).
- Ephemeren und nicht mehr bestehenden Netzen wie ehemaligen Regimen (Seibel und Raab 2003) und bankrotten Firmen (Diesner et al. 2005).
- Sehr großen Netzen, bei denen die Erhebung von personenbezogenen Daten innerhalb der Netzwerkgrenzen zu ressourcenintensiv ist, z.B. bei Studien zu den Mitgliedern der verschiedenen Gemeinschaften oder Ethnien in einer Region (Burt und Lin 1977).
- Netzen, denen keine bereits bestehende soziale Struktur zugrunde liegt, sondern die lediglich aus den Datenspuren, die von oder in ihnen generiert werden, bestehen, wie z.B. das Internet und Blogs (Adar und Adamic 2005). Wir bezeichnen solche Netze als WYSIWII - what you see is what it is (Diesner und Carley 2009a).
- Sozialen Netzen, die mit semantischen Netzen kombiniert werden (Diesner und Carley 2005).

In den genannten Fällen, die sich überlappen können, sind Textdaten wie z.B. gerichtliche Dokumente, Jahresberichte, Bücher, Nachrichtenmeldungen, Interviews, E-Mails und Webseiten oft die einzige Informationsquelle von oder über ein System. Wenn aus solchen Texten zu den hier genannten Arten von Systemen relationale Daten extrahiert werden, ist deren Validierung, also der Abgleich von erhobener und tatsächlicher Struktur, schwierig bis unmöglich. Daher sind das stringente und umfassende Testen der Extraktionsverfahren seitens der Entwickler, die Kommunikation und Rezeption dieser Ergebnisse, sowie die informierte Anwendung entsprechender Methoden, Maßzahlen und Software seitens der Nutzer unabdingbar. Dieser Beitrag soll diesem Zweck dienlich sein.

## 3 Verfahren zur Extraktion von Relationen

Im Allgemeinen erfordern Relationsextraktionsverfahren das Durchführen der folgenden Schritte: Zunächst sind ein Problem, eine Forschungsfrage oder ein konkretes Ziel, für deren Bearbeitung relationale Daten hilfreich sind, zu formulieren. Falls nicht bereits vorhanden, ist ein Korpus von Textdaten zu erheben. Entsprechende Daten fallen häufig als Nebenprodukt von Organisations- und Kommunikationsprozessen an. Ein Beispiel hierfür sind die Antworten auf offene Fragen in Interviews, die bei Datenanalysen häufig unberücksichtigt bleiben oder in kleiner Menge qualitativ ausgewertet werden. Anschließend werden in den zu untersuchenden Textdaten die relevanten Knoten und deren Verbindungen identifiziert. Darüber hinaus ermöglichen fortgeschrittene Verfahren das Klassifizieren von Knoten (multi-modale Netze) und Kanten (multirelationale Netze) gemäß vordefinierter oder aus den Daten abgeleiteter Ontologien oder Taxonomien. Knoten in semantischen Netzen werden auch als Konzepte bezeichnet. Konzepte bestehen aus ein oder mehreren Worten und geben Informationen aus dem Text in wortgetreuer, normalisierter, disambiguierter oder abstrahierter Form wieder. Kanten verbinden Konzepte und können, je nach Verfahren, binär oder gewichtet, benannt oder unbenannt, und gerichtet oder ungerichtet sein. Knoten und Kanten können, ebenfalls in Abhängigkeit des Verfahrens, mit Attributen und deren Werten versehen werden. Die extrahierten Daten werden als Graph, Liste oder Tabelle repräsentiert. Dieser Schritt markiert das Ende der Extraktionsphase, nicht aber des Untersuchungsprozesses: Es folgt die zielgerichtete Nutzung der Daten als Input z.B. für Datenbanken (McCallum 2005; Shapiro 1971), Visualisierungen (Hartley und Barnden 1997), Analysen sozialer Netze (Diesner und Carley 2005), Simulationen (Carley et al. 2007) und statistische Verfahren sowie Methoden des maschinellen Lernens und der künstlichen Intelligenz (McCallum 2005). Schließlich sind die Ergebnisse der Datennutzung zu interpretieren und zu validieren.

Im Einzelnen unterscheiden sich die Verfahren hinsichtlich ihrer Terminologie, Anwendungsbereiche, theoretischen Anbindung, Annahmen, Mechanismen für das Auffinden und Klassifizieren von Knoten und Kanten, Automatisierungsgrades, und Evaluierung. Wir konzentrieren uns in der folgenden Diskussion von Gruppen von Verfahren auf Ansätze zum Lokalisieren und Klassifizieren von Knoten und Kanten.

Wie kann man also vorgehen, wenn in einem Textkorpus die relevanten Knoten und Kanten möglichst systematisch, vollständig, richtig und effizient identifiziert und im Bedarfsfall klassifiziert werden sollen? Zunächst ist zu beachten, dass das Ziehen von Stichproben aus Texten problematisch ist, weil Sprache zu Ungunsten sinntragender bzw. rele-

vanter Terme schief verteilt ist (Zipf 1949). Das heißt, potentielle Knoten und Kanten sind mager und unregelmäßig über die Texte verteilt, während bedeutungsarme Begriffe dicht und häufig vorkommen. Das gleiche Prinzip gilt auch beim Erheben von Netzwerkdaten mit anderen Methoden wie z.B. Befragungen (Frank 2004): je weniger vollständig die Daten innerhalb der definierten Netzwerkgrenze erfasst werden, umso ungleich grösser ist die Chance, schwach vernetzte Knoten zu überrepräsentieren und zentrale Knoten zu unterrepräsentieren. Bei der Analyse komplexer, dynamischer und soziotechnischer – kurz alltäglicher – Systeme ist es also notwendig, die Daten über ein System in ihrem ganzen Umfang zu nutzen (Corman et al. 2002). Je nach System- und Korpusgröße, vorhandenen Ressourcen und erforderlicher Genauigkeit stehen dafür eine Reihe von Verfahren zur Verfügung.

### *3.1 Qualitative Textanalyse*

Bei der qualitativen Textanalyse übernimmt der Mensch das datengeleitete Identifizieren von relevanten Konzepten (*Kodes*), das anfängliche Auffinden und Kommentieren von Instanzen der Kodes in den Daten, und das Erstellen von benannten Links zwischen Kodes (Bernard und Ryan 1998). Zahlreiche Computerprogramme unterstützen Analysten beim systematischen und iterativen Assoziieren von Textpassagen mit Schlagworten oder Kodes (Kodieren, Indexieren), Erläutern des Kodes in Memos, Aggregieren ähnlicher Kodes zu Variablen und Ausgeben aller Segmente, die mit bestimmten Kodes versehen wurden ( Lewins und Silver 2007). Die Suchergebnisse können der Datenbasis als Material hinzugefügt werden (*system closure*, Richards 2002). Alle Objekte in der nutzerdefinierten Datenbasis (*hermeneutischen Einheit*), wie Texte, Kodes, Memos, multimediale Daten und Ergebnisse, können mit Attributen versehen und analysiert werden. Durch das manuelle Arrangieren und Verbinden von Variablen entstehen Netze, die implizite Beziehungen in den Daten explizit abbilden und der Entwicklung von Modellen und Theorien dienen (*Grounded Theory*, Glaser und Strauss 1967). Diese Netze können computergestützt visualisiert und manipuliert werden. Die qualitative Textanalyse dient der Exploration von Daten und Phänomen, dem Generieren von Hypothesen, und der Erlangung eines tiefgründigen Verständnisses basierend auf überschaubaren Datenmengen. Computerprogramme übernehmen hierbei keinerlei analytischen Aufgaben, sondern dienen lediglich als Arbeitsumgebung, während der Mensch die Verantwortung für das konsistente und zuverlässige Erkennen von Kodes trägt. Die Evaluierung der Ergebnisse erfolgt durch das Testen der Modelle mit neuen Daten, und die der Daten durch das Messen der Konsistenz in der Kodierung mehrerer Texte durch eine oder mehrere Personen (*Goldstandard*, King und Lowe 2003).

### *3.2 Listen, Regeln und Verfahren aus der Computerlinguistik*

Wie können das Lokalisieren und Klassifizieren relevanter Konzepte und Kanten erleichtert und beschleunigt werden? Eine Möglichkeit ist die Nutzung vorhandener Hilfsmittel wie Listen und Thesauren bzw. Wörterbücher. Sollen beispielsweise die Mitgliedschaften eines Landes in internationalen Organisationen erkannt werden, können diese Organisationen aus dem CIA World Factbook (Central Intelligence Agency) entnommen, als Liste repräsen-

tiert, und die Listeneinträge mit den Textdaten abgeglichen werden. Sind diese internationalen Organisationen als Akronym indexiert (z.B. WHO), helfen Thesauren bei der Aufschlüsselung von Symbolen (z.B. WHO, World Health Organisation). Listen und Thesauren dienen der Normalisierung, also der Indexierung verschiedener Schreibvarianten von Konzepten und der Zuweisung von Synonymen zu einheitlichen Schlagworten, der Auflösung von Mehrdeutigkeiten (Disambiguierung) durch Unterscheidung von Groß- und Kleinschreibung (Hoch versus hoch) oder mittels kurzer Phrasen (ein Hoch über Sachsen versus hoch über der Stadt), sowie als positive Filter (Züll und Alexa 2001). Solche Hilfsmittel bergen eine Reihe von Risiken: sie können unvollständig, veraltet und fehlerhaft sein. Zudem ist ihre Anwendung deterministisch, das heißt, nicht indexierte Terme wie Rechtschreibfehler und Wortableitungen werden in den Texten nicht gefunden.

Mehr Flexibilität bietet das Durchsuchen von Texten nach Instanzen abstrahierter Muster von Symbolen (*reguläre Ausdrücke*, Kleene 1956). Damit lassen sich beispielsweise Telefonnummern, Datumsangaben und URLs recht präzise identifizieren und klassifizieren. Der reguläre Ausdruck *[A-Z]* bewirkt z.B. die Ausgabe aller großgeschriebenen Worte. Deren Zuordnung zu Unterkategorien von Eigennamen (*named entities*) wie Personen, Organisationen und Orten kann mit Hilfe regulärer Ausdrücke jedoch nicht geleistet werden, sondern bedarf weiterer Verfahren (Diesner und Carley 2008).

Eine weitere Möglichkeit sind regelbasierte Verfahren, die beobachtete Regelmäßigkeiten im Aufbau oder der Formatierung von Texten generalisieren, und diese Regeln auf das gesamte Datenset anwenden. Carafalla et al. (2006) nutzen diesen Ansatz, um aus Tabellen, die in HTML ausgedrückt sind, Informationen, deren Bedeutung wie im Tabellenkopf angegeben, und deren Relationen zu extrahieren. Problematisch sind hierbei die deterministische Natur von Regeln, das mögliche Zutreffen von Regeln auf irrelevante Phrasen, dass zu Fehlern 1. Art bzw. α-Fehlern führt, und die Bewältigung von Ausnahmen, die nur durch extensive Erweiterung des Regelwerks oder gar nicht realisiert werden kann.

Die Computerlinguistik hält schließlich Verfahren zur Sprachverarbeitung bereit, mit denen Textdaten aufbereitet (Jurafsky und Martin 2000) und relevante Terme gefunden werden können (*feature selection*, Yang und Pedersen 1997). Zu diesen Verfahren zählen unter anderem das Segmentieren von Texten in Sätze und in Worte, das Entfernen nicht sinntragender Symbole und Worte wie Sonderzeichen und Präpositionen, das Rückführen von Wortableitungen auf deren Stammform bzw. Lexem (*Stemming*), das Bestimmen der Wortklasse wie z.B. Substantiv oder Adjektiv, das Erkennen von sinntragenden Mehrwortgruppen und Eigennamen, und das Separieren von relevanten versus irrelevanten Begriffe durch Termgewichtungsverfahren. Diese Verfahren sind meist vollautomatisiert und funktionieren regelbasiert und/ oder probabilistisch. Aufgrund der Komplexität natürlicher Sprache einschließlich ihrer Unregelmäßigkeit können computergestützte Verfahren eine Reihe von Stilmitteln und Mehrdeutigkeiten jedoch nicht immer korrekt verarbeiten. Dazu zählen Metaphern, umgangssprachliche Wendungen und Slang, und Worte mit gleicher Schreibweise und Wortklasse aber unterschiedlicher Bedeutung (*Homonyme*, z.B. *Leiter* als Kabel, Chef und Werkzeug).

Praktische Anwendungen zum Auffinden relevanter Konzepte kombinieren häufig die genannten Verfahren. So zum Beispiel in einer Serie von Systemen, die Politikwissenschaftler zum Kodieren und Auswerten internationaler Ereignisse entwickelt und eingesetzt haben (Schrodt et al. 2008). Dabei wird zunächst die syntaktische Oberflächenstruktur jedes

Satzes (Wortklasse pro Wort) automatisch ermittelt, um die Position von Verben und Nominalphrasen zu finden. Die relevanten Instanzen beider Wortklassen werden manuell gegen eine existierende Typologie von Ereignissen (Verben) und eine Liste von Akteuren (Eigennamen) abgeglichen und die Listen um fehlende Einträge erweitert. Um die Vergleichbarkeit von Studien und die Adaption an Veränderungen in der Weltpolitik zu gewährleisten, passt die Fachgemeinschaft diese Listen stetig an und integriert existierende Standards Dritter wie z.B. der Vereinten Nationen. Schließlich werden die Listen vollautomatisiert auf das gesamte Textset angewandt, um Triplets zu extrahieren, in denen Akteure über Ereignisseverbunden sind. Ein geübter Analyst kodiert fünf bis zehn solcher Triplets pro Stunde (Schrodt et al. 2008). Geeignete Software kann große Datenmengen in wesentlich kürzerer Zeit und vergleichbarer Qualität wie ein Mensch bei einer höheren Wahrscheinlichkeit von α-Fehlern bewältigen (King und Lowe 2003). Über alle Relationsextraktionsverfahren hinweg tauchen die Triplets aus Subjekt, Handlung und Objekt immer wieder als kleinste sinntragende Einheit semantischer Netze auf (siehe z.B. Berners-Lee et al. 2001; Franzosi 1989; Mohr 1998; van Cuilenburg et al. 1986). Je nach Verfahren kann diese Grundstruktur durch weitere Angaben wie Ort, Zeit und Attribute erweitert werden.

### *3.3 Distanzbasierte Verfahren*

Wenn die relevanten Konzepte in den Texten lokalisiert wurden, gilt es, sie zu verlinken. Ein weit verbreiteter, regelbasierter und deterministischer Ansatz hierfür ist das Windowing (Danowski 1993). Dabei definiert der Analyst die Weite eines Fensters (Anzahl von Worten) innerhalb einer ebenfalls vom Nutzer festgelegten Texteinheit, wie z.B. einer bestimmten Anzahl von Sätzen, Paragraphen oder dem gesamten Text. Das Fenster wird über den Text gezogen und die dabei darin gemeinsam erscheinenden Paare relevanter Konzepte in einer Adjazenzmatrix vermerkt. Die Datenstruktur der Adjazenzmatrix ist Grundlage für eine Reihe von Analyseverfahren:

Methoden zur Reduktion der Dimensionen von Daten sind ein Ansatz: Bei der multidimensionalen Skalierung werden die Werte in der Adjazenzmatrix räumlich interpretiert: Je öfter zwei Konzepte gemeinsam auftreten, umso kürzer ist ihre euklidische Distanz in einer zweidimensionalen Visualisierung (Doerfel und Barnett 1999; Osgood 1959). Mehrfachbedeutungen von Worten können damit allerdings häufig nicht klar unterschieden werden (Tversky und Gati, 1982). Clusteringverfahren, die auf die Adjazenzmatrix oder deren graphische Repräsentation angewandt werden, ermöglichen das Finden und Vergleichen von Gruppen häufig gemeinsam und selten mit anderen auftretender Datenpunkte (Smith und Humphreys 2006).

Carley (1997) wendet eine Kombination aus teilautomatisierter Sprachverarbeitung und Windowing (Map Analysis) auf Textdaten an, um kognitive Modelle (*mental models*) von Individuen und Gruppen zu extrahieren. Solche Modelle sind vereinfachende Darstellungen der Welt inklusive individueller Wahrnehmung (Johnson-Laird 2005; Rumelhart 1981). Die Modelle werden netzwerkanalytisch und mengenalgebraisch untersucht, verglichen und kombiniert, um die soziale Bedeutung und Entwicklung von Informationen zu erschließen. In der praktischen Anwendung der Methode kann sich die Bedeutungsfindung auf die Intensionen des Autors wie auch auf Texte als Indikatoren menschlichen Verhaltens beziehen (Bernard und Ryan 1998; Roberts 2000). Das Mapping von Sprache auf kognitive

Strukturen unterliegt einer Reihe von Annahmen: kognitive Modelle sind eine Repräsentationen der Organisation von Informationen im Gedächtnis und können als Netz repräsentiert werden, Sprache gibt uns einen Zugang zu diesen internen Strukturen, und individuelle Kognition beeinflusst soziales Verhalten. Tatsächlich sind die Beziehungen zwischen Sprache, deren Repräsentationen als Netzen, und deren Bedeutung nach wie vor unzureichend erforscht (Carley und Palmquist 1991).

Das Konstrukt der kognitiven Modelle ist theoretisch motiviert (Carley 1997). Die Wahl der Fenstergröße ist es nicht. Diese unterliegt der Entscheidung, den Tests und der Erfahrung des Forschers und kann zu Fehlern erster und zweiter Art (zuviel und zuwenig Links) führen, was Fehlinterpretationen der Ergebnisse nach sich ziehen kann (Carley 1997; Corman et al. 2002). Eine Steigerung der Genauigkeit und Validität der Ergebnisse ist durch die Kombination des Windowing mit anderen Methoden, die wir nachfolgend erläutern, möglich.

Unschärfen im Verständnis der Methodik und der Ergebnisse von Relationsextraktionsverfahren können zudem durch Überschneidungen in der Terminologie verschiedener Verfahren zur nicht-linearen Wissensrepräsentation entstehen. Diese Verfahren haben mitunter nichts gemeinsam außer der Tatsache, dass sie den Gegenstand oder das Ergebnis ihrer Analyse als semantisches Netz bezeichnen. Wir möchten ein paar dieser Begriffe unterscheiden: Mind Maps (Buzan 1984) und Concept Maps (Novak und Cañas 2008) sind graphische Darstellungen der Fakten oder persönlichen Gedanken zu einem Thema. Bei der manuellen oder computergestützten Erstellung dieser Maps definieren Personen Konzepte, die sie mit selbst benannten Kanten nach persönlichem Ermessen verbinden . Solche heuristischen Denkwerkzeuge werden zum Brainstorming und als Lernhilfe eingesetzt, stoßen aber bei umfangreichen und komplexen Themen an die Grenzen der menschlichen Wahrnehmung und Informationsverarbeitung (Hartley und Barnden 1997). Beim Semantischen Web ist der Name Programm: Menschen nutzen eine vorgegebene Beschreibungssprache, um Begriffe und deren Relationen zu definieren und damit Webinhalte für Computer interpretierbar und verwertbar zu machen (Berners-Lee et al. 2001).

### *3.4 Linguistische und kognitionswissenschaftliche Verfahren*

Bei Satzanalysen (parsing) werden die Beziehungen zwischen Worten durch die Anwendung von Grammatiken ermitteln. Grammatiken sind Regelwerke, die spezifizieren, welche Wortfolgen in einer Sprache zulässig sind. Grammatiken werden eingesetzt, um die syntaktischen Beziehungen zwischen Paaren einzelner Worte (*Dependenzgrammatik*, Tesnière 1959) oder Bestandteilen von Sätzen, die auch Mehrwortgruppen sein können *(generative Transformationsgrammatiken, kontextfreie Grammatiken*, Chomsky 1956), als hierarchische Struktur bzw. Ableitungsbaum abzubilden. Für Relationsextraktionen sind Ableitungsbäume zunächst ungeeignet, da viele Satzbestandteile für semantische Analysen irrelevant sind (magere Daten) und der Platz relevanter Worte im Baum variiert (Roberts 2000). Präzise Kenntnisse über die Wortklasse und Abhängigkeiten aller Satzbestandteile, wie sie syntaktische Satzanalysen bereitstellen, sind jedoch als Input für semantische Grammatiken essentiell (Franzosi 1989; Roberts 1997). In semantischen Grammatiken sind die Regeln und Satzbestandteile auf die Konzepte und Relationen einer bestimmten Domain abgestimmt. Den Einbußen an Generalität steht somit die Gewinnung der bedeutsa-

men Interpretierbarkeit der Ergebnisse gegenüber. Frühe semantische Grammatiken sind kognitionswissenschaftlich inspiriert. Fillmore's (1968) Kasusgrammatik geht beispielsweise davon aus, dass das Verb der zentrale Bestandteil eines Satzes ist und es zudem einen Tiefenkasus hat, der bestimmte andere Satzbestandteile zwingend oder möglicherweise erforderlich macht. Semantische Kasus, die auch Fälle oder Rollen genannt werden, kann man analog zu grammatischen Fällen, wie z.B. dem Nominativ oder Genitiv, verstehen: sie stellen Erwartungen an das Umfeld eines Wortes. So bezeichnet der *agentive* das Subjekt und der *dative* das Objekt einer Handlung. Die Instanzen der Rollen, also die Worte im Text, füllen die Positionen innerhalb eines durch die Wortidentität des Verbs vorgegebenen Rahmens (*frame*) aus. Das individuelle Füllen dieser Positionen spiegelt das Kontinuum von universellen Regeln von Sprache und stereotypen Schablonen für Situationen über kulturelle Eigenheiten bis hin zu persönlichen Erfahrungen, die nicht logisch oder korrekt sein müssen, wieder (Fillmore 1982; Minsky 1974; Woods 1975). Die individuelle Bedeutung eines Wortes in einem Netz kann somit als die Netzwerkumgebung, die durch das Wort motiviert oder aktiviert wird, erschlossen werden (Carley und Palmquist 1991; Collins und Loftus 1975). Semantische Grammatiken können erfolgreich eingesetzt werden, wenn die Grammatik den Wortschatz und die Struktur der Texte ausreichend erschöpfend abdeckt, was in der Praxis oft extensive manuelle Arbeit am Regelwerk bedeutet.

### *3.5 Ansätze aus der Künstlichen Intelligenz*

Im Gegensatz zu den bis hier vorgestellten Anwendungen von Relationsextraktionen untersucht man in der Künstlichen Intelligenz (KI) nicht, was ein relationaler Ausdruck bedeutet, sondern ob er wahr oder falsch ist (Woods 1975). In der KI werden semantische Netze erstellt, indem ein klar definierter Prozess oder ein Algorithmus manuell oder computergestützt angewandt werden, um die möglichen Bedeutungen natürlichsprachlicher Sätze in eine formale, präzise und eindeutige Repräsentation von Wissen zu übersetzen (Norvig und Russell 1995). Die KI folgt Platos Definition von Wissen als gerechtfertigten und wahren Annahmen. Auf die erhobenen relationalen Daten werden Inferenzregeln angewandt, um logische oder probabilistische Schlussfolgerungen zu ziehen, das heißt aus bestehendem Wissen neues Wissen abzuleiten. Das Verfahren wird eingesetzt, um Wissen zu managen und für Abfragen bereitzustellen (Allen und Frisch 1982), zum Aufdecken und Korrigieren von Widersprüchen in Wissensdatenbanken (Sowa 1992), zur Beweisführung, zum Verstehen natürlicher Sprache (Shapiro 1971), zur Manipulation oder dynamischen Veränderung von Netzen (Petri 1962), und zu Berechnungen auf der Grundlage graphischer Modelle, in denen Sachverständige die Richtung und Wahrscheinlichkeit der Abhängigkeiten zwischen den relevanten Variablen einer Domain darstellen (Howard 1989; Pearl 1988).

Das Inferieren von Schlüssen aus einer Wissensdatenbank ist nur sinnvoll, wenn die relationalen Ausdrücke einer bestimmten Syntax und Logik fehlerfrei und ohne Unschärfen folgen. Das Wissen einer Domain kann in einer standardisierten Beschreibungssprache (Ontologie) für gültige Konzepte und deren Beziehungen spezifiziert werden. Ontologien organisieren Konzepte in der Regel vom Allgemeinen zum Spezifischen, wobei Oberklassen ihren Unterklassen Eigenschaften vererben. Durch die Benennung der Relationen können zum Beispiel Bestandteile von Objekten und Kausalität beschrieben werden. Eine für praktische sprachverarbeitende Anwendungen einsetzbare Ontologie ist WordNet

(Fellbaum 1998). In WordNet sind bedeutungsgleiche Verben, Substantive, Adjektive und Adverbien in Synonymgruppen, sogenannten synsets, zusammengefasst. Die synsets sind mit einer kurzen Definition versehen und untereinander durch semantische Beziehungen wie Hyperonymie (Oberbegriffe) und Hyponymie (Unterbegriffe) verbunden.

Wenn die relevanten Konzepte in einem Satz identifiziert wurden, sind sie gemäß einer Logik zu verbinden: Die Aussagenlogik erlaubt die Verknüpfung von Aussagen durch Verneinung, Konjunktion, Disjunktion, Implikation und Äquivalenz. Die Prädikatenlogik behandelt Aussagen als Tuples von Objekten und Prädikaten (Relationen) und stellt Quantoren bereit, mit denen man ausdrücken kann, ob eine Aussage auf alle oder weniger als alle Objekte zutrifft (Allen und Frisch 1982; Janas und Schwind 1979).

Die Erstellung von semantischen Netzen nach den Prinzipien der KI ist eine Übersetzung von Sprache in relationale Daten, die als Netzwerkvisualisierung oder linearer Ausdruck ausgegeben werden können. Beide Ausgaben sind isomorph, das heißt sie drücken das Gleiche aus (Sowa 1992). Weil die Syntax von Visualisierungen (das Netzwerklayout) im Gegensatz zu linearen Ausdrücken in der Regel keine Entsprechung in der Logik hat, hat sich dafür keine einheitliche Lösung durchgesetzt (Hartley und Barnden 1997). Die Visualisierung kann damit nicht automatisiert analysiert werden; ein Problem, dass sich nicht auf die KI beschränkt. Weitere Nachteile ergeben sich bei der Wissensverarbeitung im Sinne der KI dadurch, dass die verwendeten Symbole und Regeln oft nicht allgemeingültig, sondern auf eine bestimmte Anwendung zugeschnitten und somit nur mit großem manuellen Aufwand wiederverwertbar, in größerem Maßstab anwendbar und auf andere Projekte übertragbar sind (Minsky 1974). Schließlich führen lokale Unstimmigkeiten zur globalen Inkonsistenz der Wissensdatenbank und damit zum Scheitern valider Schlussfolgerungen.

### *3.6 Statistische Verfahren und maschinelles Lernen*

Große Mengen von Textdaten können einfach, schnell und billig gesammelt und gespeichert werden. Die systematische, effiziente und kontrollierte Extraktion und Auswertung von Instanzen nutzerdefinierter Knoten- und Kantenklassen aus sequentiellen Daten erfordert adäquate Techniken, Software und Maße sowie deren sachkundige Anwendung. Wenn die zeitlichen, finanziellen und personellen bzw. kognitiven Ressourcen für die Verarbeitung großer Textmengen beschränkt sind und eine Stichprobenziehung ungeeignet ist, können Netzwerkdaten durch Methoden der Statistik und des maschinellen Lernens erhoben werden. Maschinelles Lernen (ML) sind Verfahren, bei denen das System seine Leistung hinsichtlich eines bestimmten Kriteriums, z.B. der Genauigkeit, auf der Grundlage von gesammelter Erfahrung eigenständig verbessert (Mitchell 1997). Dabei werden Modelle oder Klassifikatoren gelernt, die sich auf neue Daten mit abschätzbarer Genauigkeit anwenden lassen. Verfahren aus der Statistik und dem maschinellen Lernen sind zudem als Erweiterung in alle der in diesem Beitrag erwähnten Familien von Verfahren integriert worden; bislang mit Ausnahme der qualitativen Textanalyse.

Relationale Daten, die mittels Statistik und ML aus Texten gewonnen werden, spiegeln nicht die Wahrheit wieder. Vielmehr sind sie eine Annährung (Approximation) an die in den Texten enthaltenen Netzwerkdaten inklusive der Ungenauigkeiten, die natürliche Sprache in sich birgt, und der probabilistischen Natur von Statistik und ML. Die bestmögliche Approximation ist eine Herausforderung an die moderne Wissenschaft und Technik.

Die Förderung von Forschungswettbewerben, das Bereitstellen von Daten und die Entwicklung stringenter Evaluationsmaße durch die US-Regierung hat seit den 90ern zu einer Fülle von erfolgreichen Innovationen und Produkten auf dem Gebiet der Relationsextraktion geführt (Doddington et al. 2004; Grisham und Sundheim 1996; Miller et al. 2000).

Diesner und Carley (2008) haben beispielsweise solche Daten und ein maschinelles Lernverfahren genutzt, um ein Modell zu entwickeln, dass die Instanzen der Knotenklassen *Akteur*, *Organisation*, *Ereignis*, *Wissen*, *Ressource*, *Aufgabe*, *Ort*, und *Zeit* automatisiert in englischsprachigen Texten findet und der entsprechenden Knotenklasse zuweist. Das Modell ist in AutoMap, einer Software für relationale Textanalyse, verfügbar, und hat eine aus Trefferquote und Genauigkeit kombinierte Prognosegenauigkeit von 83% (ebd.). Die Lerntechnologie, mit der das Modell entwickelt wurde, kann wiederverwendet werden, um Modelle für andere, nutzerdefinierte Knotenklassen zu erstellen. Das genutzte Lernverfahren heißt Conditional Random Fields (CRF) und gehört zur Familie konditionaler, ungerichteter und graphischer Modelle (Dietterich 2002; Lafferty et al. 2001). Diese Modelle eignen sich aufgrund ihrer mathematischen Eigenschaften gut zur Arbeit mit großen Datenmengen in denen nur wenige Konzepte relevant sind (magere Daten). Zu diesen Eigenschaften gehört das Schließen von lokalem Kontext auf global optimierte Lösungen und die Berücksichtigung lexikalischer, syntaktischer und sich gegenseitig beeinflussenden Texteigenschaften (*features*), die über weite Textdistanzen zum Tragen kommen (Bunescu und Mooney 2007; Culotta et al. 2006).

Nachdem die relevanten Konzepte identifiziert wurden, helfen statistische oder ML Verfahren auch bei deren Verlinkung. Dazu einige Beispiele: Zelenko et al. (2003) nutzen mehrdimensionale Ähnlichkeitsfunktionen (*kernels*), um potenzielle und mit Wortklassen annotierte Relationen aus neuen Texten mit einem Modell bereits bestätigter Relationen zu vergleichen. Brin's System (1999) sucht in großen Textmengen wie z.B. dem Internet nach Übereinstimmung mit einem kleinen Set von abstrahierten Repräsentationen einiger weniger Relationen. Aus den gefundenen Übereinstimmung werden die an vielen der Relationen beteiligten Terme herausgelöst, um dann nach weiteren Mustern, in denen diese Terme hinreichend oft vorkommen, zu suchen (*bootstrapping*). Dieser Zyklus wird wiederholt bis die Basis an Entsprechungen, also die eigentlichen Relationen, und Mustern ausreichend groß ist. Culotta et al. (2006) extrahieren Kanten mittels CRF und Speichern diese Kanten in einer Datenbank. Dort werden weitere Data Mining Verfahren auf die Relationen angewandt, um diese zu bereinigen und neue, implizite Beziehungen aufzudecken.

Im Gegensatz zu den hierin besprochenen KI Verfahren ist die Überführung von Texten in relationale Daten mittels statistischer und ML Verfahren keine Übersetzung, sondern eine Transformation. Das Endprodukt dieser Transformation hat keine direkte Entsprechung in den Originaldaten, sondern ist das Ergebnis von strukturbewahrenden- und enthüllenden Reduktions- und Abstraktionsprozessen, die es uns ermöglichen, die Mechanismen und Dynamiken von Netzen klarer zu sehen und zu kommunizieren (Franzosi 1989; McCallum 2005; Mohr 1998). "Yet only through such abstractions can we come to understand the structural interrelations among the confusing mass of concrete events" (White 1993: 159).

Die Lösungen, die approximative Verfahren vorschlagen, sind nicht unbedingt die richtigen, sondern die wahrscheinlichsten. Deshalb ist die bestmögliche Approximation auch eine soziale Herausforderung: Die entsprechenden Verfahren sind komplex hinsichtlich ihrer Annahmen, Algorithmen und Parameter, und fordern daher den Entwicklern eine

Reihe von Entscheidungen zwischen möglichen Alternativen ab. Diese Entscheidungen können das Verhalten der Verfahren, Maßzahlen und Produkte, welche die Entwickler den Endnutzer zur Verfügung stellen, beeinflussen (Diesner und Carley 2009b). Das ist dann kritisch, wenn ein Teil der Varianz im Netzwerk nicht durch das zugrundeliegende soziale System, sondern das Analysewerkzeug induziert wird. Wir argumentieren, dass deshalb mehr Brücken zwischen beiden Seiten benötigt werden: Entwickler sollten ihre Lösungen mit Sachverstand und Sorgfalt erstellen, deren Robustheit und Verhalten stringent testen und die Testergebnisse klar verständlich kommunizieren. Anwender sollten sich bemühen, die Produkte und Methoden, die sie nutzen, hinsichtlich ihrer Annahmen und ihres Verhaltens zu verstehen und Ergebnisse dementsprechend zu interpretieren. Dazu gehört zum Beispiel die Kenntnis darüber, ob eine Methode deterministisch oder probabilistisch ist sowie die adäquate Auswertung der Resultate. Diese zusätzlichen Anstrengungen aller involvierten Partner sind langfristig für eine aussagekräftige Nutzung der Netzwerkanalyse essenziell.

## 4 Schlussfolgerung und Ausblick

Die Extraktion relationaler Daten aus Texten ist eine interdisziplinäre Methodik, die Komponenten aus verschiedenen Disziplinen integriert und Paradigmen wie quantitative und qualitative Forschung zusammenführt. Damit wird es möglich, explizit und implizit in Texten enthaltene Konzepte und deren Verbindungen aufzudecken und als Input für weitere Zwecke bereitzustellen. Herkömmliche Textanalyseverfahren ermöglichen in der Regel die tiefgründige Analyse kleiner Textmengen oder die oberflächige Analyse großer Datensätze (Corman et al. 2002). Um jedoch große Textmengen zu bearbeiten und Theorien über komplexe, soziotechnische Systeme zu überprüfen, benötigt man automatisierte und skalierbare Verfahren, die ein tiefgründiges Verständnis der Interaktionen und Wechselwirkungen zwischen den Elementen dieser Systeme ermöglichen. Verfahren des maschinellen Lernens, die Bausteine aus weiteren Disziplinen wie den Sprach-, Geistes- und Sozialwissenschaften sachkundig auswählen und integrieren, sind dafür möglicherweise der vielversprechendste Ansatz (Diesner und Carley 2008; McCallum 2005; Van Atteveldt 2008). Dieser Weg stellt Nutzer wie Entwickler vor Herausforderungen, zu deren Bewältigung Kommunikation genauso wichtig ist wie das öffentlich zugängliche Bereitstellen, Formalisieren und Integrieren von Daten, Prozessen, Algorithmen und Ein- und Ausgabeformaten.

## 5 Literatur

*Adar, Eytan* und *Lada A. Adamic*, 2005: Tracking Information Epidemics in Blogspace. Proc. of IEEE/WIC/ACM International Conference on Web Intelligence, September 2005, Compiegne, Frankreich: 207-214.

*Allen, James F.* und *Allen M. Frisch*, 1982: What's in a semantic network? Proc. of 20th annual meeting of Association for Computational Linguistics Toronto, Canada: 19-27.

*Baker, Wayne E.* und *Robert R. Faulkner*, 1993: The Social Organization of Conspiracy: Illegal Networks in the Heavy Electrical Equipment Industry. American Sociological Review 58(6): 837-860.

*Berelson, Bernard*, 1952: Content analysis in communication research. Glencoe, Ill: Free Press.

*Bernard, H. Russel* und *Gery W. Ryan*, 1998: Text analysis: Qualitative and quantitative methods. S. 595-646 in: *H. Russel Bernard* (Hg.), Handbook of methods in cultural anthropology, Walnut Creek: Altamira Press.

*Berners-Lee, Tim*, *James Hendler* und *Ora Lassila*, 2001: The Semantic Web. Scientific American 284(5): 34-43.

*Brin, Sergey*, 1999: Extracting Patterns and Relations from the World Wide Web. WebDB Workshop at 6th International Conference on Extending Database Technology (EDBT), März 1998, Valencia, Spanien: 172–183.

*Bunescu, Razvan* und *Raymond J. Mooney*, 2007: Statistical Relational Learning for Natural Language Information Extraction. S. 535-552 in: *Lise Getoor* und *Ben Taskar* (Hg.), Statistical Relational Learning. Cambridge: MIT Press.

*Burt, Ronald* und *Nan Lin*, 1977: Network Time Series from Archival Records. S. 224-254 in: *David R. Heise* (Hg.), Sociological Methodology, San Francisco, CA: Jossey-Bass.

*Buzan, Tony*, 1984: Make the Most of Your Mind. New York, NY: Simon and Schuster.

*Cafarella, Michael J.*, *Michele Banko* und *Oren Etzioni*, 2006: Relational web search. Proc. of World Wide Web Conference (WWW), Mai 2006, Edinburgh, UK.

*Carley, Kathleen M.*, 1997: Network text analysis: The network position of concepts. S. 79-100 in: *Carl W. Roberts* (Hg.), Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts. Mahwah, NJ: Lawrence Erlbaum Associates.

*Carley, Kathleen M.*, *Jana Diesner*, *Jeffrey Reminga* und *Maksim Tsvetovat*, 2007: Toward an interoperable dynamic network analysis toolkit. Decision Support Systems. 43(3): 1324-1347.

*Carley, Kathleen M.* und *Michael Palmquist*, 1991: Extracting, Representing, and Analyzing Mental Models. Social Forces 70(3): 601 - 636.

*Central Intelligence Agency*. World Factbook: Available from: https://www.cia.gov/library/publications/the-world-factbook/.

*Chomsky, Noam*, 1956: Three models for the description of language. IRE Transactions on Information Theory 2(3): 113-124.

*Collins, Allan M.* und *Elisabeth F. Loftus*, 1975: A spreading-activation theory of semantic processing. Psychological Review 82: 407-428.

*Corman, Stephen R.*, *Timothy Kuhn*, *Robert D. McPhee*, und *Kevin J. Dooley*, 2002: Studying Complex Discursive Systems: Centering Resonance Analysis of Communication. Human Communication Research 28: 157-206.

*Culotta, Aron, Andrew McCallum und Jonathan Betz*, 2006: Integrating probabilistic extraction models and data mining to discover relations and patterns in text. Proc. Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL), Juni 2006, New York, NY.

*Danowski, James A.*, 1993: Network Analysis of Message Content. Progress in Communication Sciences 12: 198-221.

*Diesner, Jana* und *Kathleen M. Carley*, 2005: Revealing Social Structure from Texts: Meta-Matrix Text Analysis as a novel method for Network Text Analysis. S. 81-108 in: *V. K. Narayanan* and *Deborah J. Armstrong* (Hg.), Causal Mapping for Information Systems and Technology Research: Approaches, Advances, and Illustrations, Harrisburg, PA: Idea Group Publishing.

*Diesner, Jana* und *Kathleen M. Carley*, 2008: Conditional Random Fields for Entity Extraction and Ontological Text Coding. Journal of Computational and Mathematical Organization Theory 14(3): 248-262.

*Diesner, Jana* und *Kathleen M. Carley*, 2009a: WYSIWII - What You See Is What It Is: Informed Approximation of Relational Data from Texts. Presentation General Online Research (GOR), April 2009, Wien, Österreich.

*Diesner, Jana* und *Kathleen M. Carley* 2009b. He says, she says. Pat says, Tricia says. How much reference resolution matters for entity extraction, relation extraction, and social network analysis. Proceedings of IEEE Symposium on Computational Intelligence for Security and Defence Applications (CISDA), Juli 2009, Ottawa, Canada.

*Diesner, Jana*, *Kathleen M. Carley* und *Harald Katzmair*, 2007: The morphology of a breakdown. How the semantics and mechanics of communication networks from an organization in crises relate. Präsentation, XXVII Sunbelt Social Network Conference, Mai 2007, Korfu, Griechenland.

*Diesner, Jana*, *Terrill L. Frantz* und *Kathleen M. Carley*, 2005: Communication Networks from the Enron Email Corpus "It's Always About the People. Enron is no Different". Journal of Computational and Mathematical Organization 11(3): 201-228.

*Dietterich, Thomas G.*, 2002: Machine Learning for Sequential Data: A Review. Proc. of Joint IAPR International Workshops SSPR 2002 and SPR 2002, August 2002, Windsor, ON, Canada: 15-33.

*Doddington, George*, *Alexis Mitchell*, *Mark Przybocki*, *Lance Ramshaw*, *Stephanie Strassel* und *Ralph Weischedel*, 2004: The Automatic Content Extraction (ACE) Program–Tasks, Data, and Evaluation. Proc. of Language Resources and Evaluation Conference (LREC), Mai 2004, Lissabon, Portugal: 837-840.

*Doerfel, Marya*, 1998: What Constitutes Semantic Network Analysis? A Comparison of Research and Methodologies. Connections 21(2): 16-26.

*Doerfel, Marya* und *George A. Barnett*, 1999: A Semantic Network Analysis of the International Communication Association. Human Communication Research 25(4): 589-603.

*Fellbaum, Christiane*, 1998: WordNet: An electronic lexical database. Cambridge MA: MIT Press.

*Fillmore, Charles J.*, 1982: Frame Semantics. S. 111-137 in: *The Linguistic Society of Korea* (Hg.), Linguistics in the morning calm. Seoul, Süd Korea: Hanshin Publishing Co.

*Fillmore, Charles J.*, 1968: The Case for Case. S. 1-88 in: *Emon Bach* and *Robert T. Harms* (Hg.), Universals in Linguistic Theory. New York: Holt, Rinehart and Winston.

*Frank, Ove*, 2004: Network sampling and model fitting. S. 31–56 in: *Peter J. Carrington*, *John Scott* und *Stanley Wasserman* (Hg.), Models and methods in social network analysis. New York: Cambridge University Press.

*Franzosi, Roberto*, 1989: From words to numbers: A generalized and linguistics-based coding procedure for collecting textual data. Sociological Methodology 19: 225-257.

*Gerner, Deborah*, *Phillip A. Schrodt, Ronald A. Francisco* und *Judith L. Weddle*, 1994: Machine Coding of Event Data Using Regional and International Sources. International Studies Quarterly 38(1): 91-119.

*Glaser, B.* und *A. Strauss*, 1967: The Discovery of Grounded Theory: Strategies for Qualitative Research. New York, NY: Aldine.

*Grisham, Ralph* und *Beth Sundheim*, 1996: Message understanding conference - 6: A brief history. Proc. of 16th International Conference on Computational Linguistics, Kopenhagen, Dänemark, Juni 1996.

*Hartley, Roger* und *John Barnden*, 1997: Semantic networks: visualizations of knowledge. Trends in Cognitive Sciences 1(5): 169-175.

*Howard, Ronald A.*, 1989: Knowledge maps. Management Science 35(8): 903-922.

*Janas, Jtirgen und Camilla Schwind*, 1979: Extensional Semantic Networks. S. 267-302 in: *Nicholas V. Findler* (Hg.), Associative Networks. Representation and Use of Knowledge by Computers. New York u.a.: Academic Press.

*Johnson-Laird, Phil N.*, 2005: The history of mental models. S. 179–212 in: *Ken Manktelow* und *Man C. Chung* (Hg.), Psychology of Reasoning: Theoretical and Historical Perspectives. London: Psychology Press.

*Jurafsky, Daniel* und *James H. Martin*, 2000: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Upper Sadle River NJ: Prentice Hall.

*King, Gary* und *Will Lowe*, 2003: An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design. International Organization 57(3): 617-642.

*Kleene, Stephen*, 1956: Representation of events in nerve nets and finite automata. S. 3-41 in: *Claude Shannon* und *John McCarthy* (Hg.), Automata Studies. Princeton NJ: Princeton University Press.

*Kleinberg, Jon*, 2003: Bursty and Hierarchical Structure in Streams. Data Mining and Knowledge Discovery 7(4): 373-397.

*Krackhardt, David*, 1987: Cognitive social structures. Social Networks 9: 109-134.

*Krebs, Valdis E.*, 2002: Mapping networks of terrorist cells. Connections 24(3): 43-52.

*Krippendorff, Klaus*, 2004: Content analysis: An introduction to its methodology. Thousand Oaks CA: Sage.

*Lafferty, John*, *Andrew McCallum* und *Fernando Pereira*, 2001: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proc. of 18th International Conference on Machine Learning, Juni 2001, Willliamstown, MA: 282-289.

*Lewins, Ann* und *Christina Silver*, 2007: Using software in qualitative research: a step-by-step guide. London: Sage.

*McCallum, Andrew*, 2005: Information extraction: distilling structured data from unstructured text. ACM Queue 3(9): 48-57.

*Miller, Scott*, *Heidi Fox*, *Lance Ramshaw* und *Ralph Weischedel*, 2000: A novel use of statistical parsing to extract information from text. Proc. of 1st Conference of North American chapter of the Association for Computational Linguistics (NAACL), Seattle, WA: 226-233.

*Minsky, Marvin*, 1974: A Framework for Representing Knowledge. MIT-AI Laboratory Memo 306.

*Mitchell, Tom*, 1997: Machine Learning. Muggleton: McGraw-Hill.

*Mohr, John W.*, 1998: Measuring Meaning Structures. Annual Review of Sociology 24(1): 345-370.

*Norvig, Peter* und *Stuart Russell*, 1995: Artificial Intelligence: A Modern Approach. Upper Saddle River: Pearson Education.

*Novak, Joseph D.* und *Alberto Cañas*, 2008: The Theory Underlying Concept Maps and How to Construct Them. Florida Institute for Human and Machine Cognition, Report No. IHMC CmapTools Rev 01-2008.

*Osgood, Charles E.*, 1959: The representational model and relevant research methods. S. 33-88 in: *Ithiel de Sola Pool* (Hg.), Trends in content analysis. Urbana, IL: University of Illinois Press.

*Pearl, Judea*, 1988: Probabilistic reasoning in intelligent systems: networks of plausible inference. San Francisco: Morgan Kaufmann.

*Petri, Carl Adam*, 1962: Kommunikation mit Automaten. Universität Bonn, Ph. D. Dissertationsschrift.

*Richards, Tom*, 2002: An intellectual history of NUD* IST and NVivo. International Journal of Social Research Methodology 5(3): 199-214.

*Roberts, Carl W.*, 1997: A Generic Semantic Grammar for Quantitative Text Analysis: Applications to East and West Berlin Radio News Content from 1979. Sociological Methodology 27: 89-129.

*Roberts, Carl W.*, 2000: A Conceptual Framework for Quantitative Text Analysis. Quality and Quantity 34(3): 259-274.

*Rumelhart, David E.*, 1981: Schemata: The building blocks of cognition. Comprehension and teaching: Research reviews: 3-26.

*Schrodt, Phillip A.*, *Ömür Yilmaz*, *Deborah J. Gerner* und *Dennis Hermick*, 2008: Coding Sub-State Actors using the CAMEO (Conflict and Mediation Event Observations) Actor Coding Framework. Präsentation, Annual Meeting of the International Studies Association, März 2008, San Francisco, CA.

*Seibel, Wolfgang* und *Jörg Raab*, 2003: Verfolgungsnetzwerke. Kölner Zeitschrift für Soziologie und Sozialpsychologie 55(2): 197-230.

*Shapiro, Stuart C.*, 1971: A net structure for semantic information storage, deduction and retrieval. Proc. of Second International Joint Conference on Artificial Intelligence: 512-523.

*Smith, Andrew E.* und *Michael S. Humphreys*, 2006: Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping. Behavior Research Methods 38(2): 262-279.

*Sowa, John F.*, 1992: Semantic Networks. S. 1493-1511 in: *Stuart C. Shapiro* (Hg.), Encyclopedia of Artificial Intelligence. New York: Wiley and Sons.

*Tesnière, Lucien*, 1959: Elements de syntaxestructurale. Paris: Klincksieck.

*Tversky, Amos,* und *Itamar Gati*, 1982: Similarity, separability, and the triangle inequality. Psychological Review, 89(2): 123-154.

*Van Atteveldt, Wouter*, 2008: Semantic network analysis: Techniques for extracting, representing, and querying media content. Charleston: Book Surge Publishers.

*van Cuilenburg, Jan J.*, *Jan Kleinnijenhuis* und *Jan A. de Ridder*, 1986: A Theory of Evaluative Discourse: Towards a Graph Theory of Journalistic Texts. European Journal of Communication 1(1): 65-96.

*White, Harrison C.*, 1993: Canvases and careers: institutional change in the French painting world. Chicago: University of Chicago Press.

*Wiebe, Janyce M.*, 2000: Learning Subjective Adjectives from Corpora. Proc. of 17th National Conference on Artificial Intelligence (AAAI) 2000, Juli 2000, Austin, TX: 735-741.

*Woods, William A.*, 1975: What's in a link: Foundations for semantic networks. S. 35-82 in: *Daniel G. Bobrow* und *Allan Collins* (Hg.), Representation and Understanding: Studies in Cognitive Science. New York: Academic Press.

*Yang, Yiming* und *Jan O. Pedersen*, 1997: A comparative study on feature selection in text categorization. Proc. 14th International Conference on Machine Learning (ICML), Nashville, TN.

*Zelenko, Dmitry*, *Chinatsu Aone* und *Anthony Richardella*, 2003: Kernel methods for relation extraction. Journal of Machine Learning Research 3(2): 1083-1106.

*Zipf, George K.*, 1949: Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Cambridge, MA: Addison-Wesley Press.

*Züll, Cornelia* und *Melina Alexa*, 2001: Automatisches Codieren von Textdaten. Ein Überblick über neue Entwicklungen. S. 303-317 in: *Werner Wirth* und *Edmund Lauf* (Hg.), Inhaltsanalyse - Perspektiven, Probleme, Potenziale. Köln: Herbert von Halem.

Diese Publikation wurde unter anderem gefördert von: National Science Foundation (IGERT Programm DGE-9972762), Office of Naval Research (ONR, MMV & ROE N00014-06-1-0104), Army Research Institute (W91WAW07C0063), Army Research Lab (DAAD19-01-2-0009), AFOSR GMU MURI (FA9550-05-1-0388), und ONR MURI (N000140811186). Die hierin enthaltenen Ansichten und Schlussfolgerungen sind die der Autoren und sollten weder als explizit noch implizit repräsentativ für offizielle Grundsätze und Richtlinien der Sponsoren und der U.S. Regierung interpretiert werden.