Kognitive Interpretation der lexikalischen Relationen und ihre Anwendung auf die Indexierung

(Theoretische Untersuchung und Implementierung)

Inaugural-Dissertation zur Erlangung der Doktorwürde
der
Philosophischen Fakultät
der
Universität zu Köln

Vorgelegt von
Ick-Su Sohn
aus Daegu, Südkorea

1. Berichterstatter:
   Prof. Dr. Manfred Thaller
2. Berichterstatter:
   Prof. Dr. Jürgen Rolshoven

Tag der Disputation: 11. Juli 2008

INHALTVERZEICHNIS

1. Einleitung

1.1 Problemstellung und Ziel der Untersuchung

Indexierung bezeichnet den Vorgang der Erstellung eines Index, der als die Kernkomponente in einem Information Retrieval System (= IRS) gilt[1] und über dessen Qualität entscheidet (vgl. Bekavac 2001:18, Cleveland/Cleveland 2001:97). Der Hauptzweck der Indexierung besteht in der Repräsentation und dem Retrieval der Dokumente, indem die inhaltlich wichtigen Informationen für das Retrieval bestimmt werden (vgl. Langcaster 1998:1, Mani 2001:23). Für die Indexierung ist die Ermittlung der konzeptuellen Struktur eines Dokumentes erforderlich (vgl. Langcaster 1998:8 ff.).[2] Obwohl die maschinelle Umsetzung dieser Tätigkeit dem Aufgabenbereich der künstlichen Intelligenz (KI) bzw. der Computerlinguistik zugehört, ist die Rolle dieser Disziplinen bei der automatischen Indexierung noch sehr bescheiden.[3] Das lässt sich auf drei wissenschaftliche Probleme zurückführen, die noch unverstanden bzw. nicht genug erforscht sind (vgl. Chowdhury 1999:333 f.): Erstens die kognitiven Prozesse, die für die prozeduralen Aspekte des Sprachverstehens zuständig sind. Zweitens die Frage, was Bedeutung eigentlich ist und wie man sie objektiv ermitteln kann (das sogennante Bedeutungsproblem). Und drittens die Frage nach der Ermittlung und der Repräsentation des Weltwissens, das alles Wissen umfasst, das keinen speziellen Wissensarten wie sprachlichem Wissen oder Handlungswissen zugeordnet werden kann, aber für das Sprachverstehen nötig ist. Dies führt dazu, dass computerlinguistische Methoden im Rahmen der Indexierung bzw. Inhaltsanalyse meistens sehr eingeschränkt eingesetzt werden. Forschungsergebnisse besagen hierbei, dass morphologische Verfahren wie

[1] Der Begriff „Information Retrieval“ wird in dieser Arbeit als Synonym von „Text Retrieval“ verwendet.

[2] Zur Bezeichnung dieser Aufgabe werden in der Bibliothekswissenschaft oft Termini wie „Inhaltsanalyse“ oder Inhaltserschließung“ verwendet (vgl. Bertram 2005, Langridge 1994, Salton 1988:439).

[3] Zum Beispiel haben Frants/Shapiro/Voiskunski (1997:155) dargestellt, warum besonders die Linguistik bei der Indexierung kaum berücksichtigt wird und hauptsächlich statistik-basierte Methoden bevorzugt werden: „When considering the main directions in automatic indexing, we have already mentioned the problem of determining the most important (for the meaning) terms of documents. One would say that certain linguistic method (algorithms) should be used for this purpose first. However, linguists lack any successful algorithms. That is why researchers use mainly statistical algorithms that calculate the measure of meaning in a term of the document. This seems very convenient for the methods incorporating descriptor weights, because the measures of meaning calculated can be considered as a weight.“

Grundformerkennung und Analyse der Komposita bessere Retrievalergebnisse ermöglichen, syntaktische Verfahren aber weniger geeignet sind (vgl. Krause 1996:86 f., Nohr 2001:62-65). Ein Beispiel für den Einsatz der computerlinguistischen Methoden auf der semantischen Ebene ist die Anwendung des Konzepts der lexikalischen Kohäsion (vgl. Halliday/Hasan 1976:274-292), die als Textkonstitution durch lexikalische Relationen wie Hyponymie, Hyperonymie, Meronymie, Synonymie etc. definiert ist. Als Beispiele für lexikalische Kohäsion finden sich im folgenden Text Hyponymien von *bear* zu *animal* und von *biologist* zu *scientist* sowie die Wiederholung (Rekurrenz) von *bear*:

> „Many wild **bears** have become ‚garbage junkies', feeding from dumps around human developments. To avoid potentially dangerous clashes between them and humans**, scientists** are trying to rehabilitate the **animals** by drugging them and releasing them in uninhabited areas. Although some **biologists** deny that the mind-altering drug was responsible for uncharacteristic behaviour of this particular **bears**, no research has been done into the effects of giving grizzly **bears** *or* other mammals repeated doses of phencyclidine." (Hoey 1991:40)

Der entscheidende Vorteil, den die Analyse der lexikalischen Kohäsion für die maschinelle Textanalyse bietet, besteht darin, dass die im Text kodierten semantischen Informationen auf diese Weise im Vergleich zu anderen, auf künstlicher Intelligenz basierten Methoden, die normalerweise eine sehr aufwändige Wissensbasis und Prozesszeit benötigen und dadurch selten in der Praxis einsetzbar sind, (vgl. Nohr 2001:70) relativ effizient aufgefunden werden können. Voorhees (1998) verwendet das Konzept der lexikalischen Kohäsion für die Auflösung von Polysemien im IRS, weil er davon überzeugt ist, dass ein auf Wortbedeutung basierter Matching-Vorgang eine bessere Retrievalqualität bietet. Weiter berechnen Stairmand (1997) und Stairmand/Black (1997) anhand der lexikalischen Kohäsion die Relevanz des Kontexts, um anschließend die ermittelten Werte den Termen, die den betreffenden Kontext ausmachen, als Termgewichtung zuzuweisen. Die Ergebnisse beider Arbeiten stimmen nicht ganz miteinander überein. Während Voorhees (1998) meint, dass die Disambiguierung der Wörter durch Hyperonymie[4] nicht sehr gut funktioniert und die Retrievalqualität daher nicht unbedingt davon profitieren kann, zeigen Stair-

[4] Das Disambiguierungsverfahren von Voorhees (1998:289-292) basiert auf der Verkettung durch Hyperonymie/Hyponymie. Die Belegform eines Wortes, die die meisten Verbindungen zu den Belegformen der anderen Wörter hat, wird als passende Belegform bestimmt.

mand und Black, dass sich durch Disambiguierung[5] die Präzision der Retrievalergebnisse im Vergleich zum SMART-IRS (Salton/Buckley 1988) steigert. Die Gemeinsamkeit beider Arbeiten liegt darin, dass die lexikalische Kohäsion hauptsächlich zur Auffindung von lexikalischen Ketten[6] verwendet wird, aus denen grobe kontextuelle Informationen gewonnen werden können, die den ganzen Text oder Textteile repräsentieren.

Die lexikalische Kohäsion wird auch in dem Indexierungsverfahren, das in dieser Arbeit vorgestellt wird, eine zentrale Rolle spielen, wobei die Berücksichtigung der kognitiven Eigenschaften der einzelnen lexikalischen Relationen im Mittelpunkt steht. Die kognitive Interpretation der lexikalischen Kohäsion und ihre Anwendung auf die Indexierung beruhen auf dem klassischen symbolverarbeitenden Ansatz in der Kognitionswissenschaft,[7] also auf der Annahme, dass Sprache als Symbolsystem einen direkten Zugang zur kognitiven Ebene ermöglicht und die Struktur der Kognition daher durch eine Analyse der sprachlichen Struktur ermittelt werden kann (vgl. Anderson 1995a:361, Hellwig 1984:71). Ziel dieser Arbeit ist es zu untersuchen, welche kognitiven Eigenschaften der den Texten zugrunde liegenden konzeptuellen Strukturen durch die lexikalischen Relationen widergespiegelt werden, und die aus dieser Analyse gewonnenen Erkenntnisse zur Erhöhung der Indexqualität zu verwenden.

[5] Die Erkennung der lexikalischen Kohäsion bei Stairmand und Black unterscheidet sich von Voorhees hauptsächlich durch zwei Faktoren: Erstens berücksichtigen sie nicht nur Hyperonymie/Hyponymie, sondern auch andere lexikalische Relationen wie Wiederholung, Synonymie und Antonymie. Zweitens ziehen sie die Distanz der Wörter im Text in Betracht, die gemeinsam eine Kohäsionskette bilden (vgl. Stokes 2004: 34 f. in Bezugnahme auf Stairmand/Black 1997).

[6] Man kann zwar bereits ein Vorkommen der lexikalischen Relation schon als lexikalische Kette ansehen, aber eine typische lexikalische Kette schließt mehrere lexikalische Relationen wie *bears*–- a*nimals*–- *bears* im obigen Beispieltext mit ein.

[7] Ein weiterer Ansatz der kognitiven Forschung ist das Konzept der sog. „neuronalen Netze“ (vgl. Schnotz 1994:119-142, Anderson 1995a:15 f.). Shastrie (1988) hat ein sog. „hybrides Modell“ vorgestellt, das beide Ansätze kombiniert.

1.2 Methode und Aufbau der Untersuchung

Indexierung setzt die Analyse der konzeptuellen Struktur eines Textdokuments voraus (vgl. Hjørland 1997:41, Langcaster 1997:8-14), wenn man den kognitiven Verstehensprozess abbilden will. Textverstehen ist nämlich nur möglich, wenn der Rezipient entsprechende Konzepte bzw. Begriffe gebildet hat:

> „Die Entwicklung der relevanten und geeigneten Begriffe ist notwendige Voraussetzung für jeden Verstehensakt. […] Umgekehrt ist die Entwicklung, d. h. die Veränderung und Erweiterung von Begriffen, eine direkte Folge aufeinanderfolgender Verstehensakte." (Seiler 1984:64)

Konzepte, die als Wissenselemente gelten (vgl. Weinert/Waldman 1988:162), sind kognitive Einheiten, die Wissen über Abschnitte unserer Umwelt repräsentieren (vgl. Hoffmann 1986:56, Konerding 1993:88 f.). Einem Konzept kann ein Zeichen zugeordnet werden oder nicht (Seiler 1984:9). Für die Repräsentation von Satzbedeutung und Textbedeutung wurde aus der formalen Semantik der Begriff „Proposition" übernommen (vgl. van Dijk 1980:25 f., van Dijk/Kintsch 1983:37-41, Engelkamp 1984a:36 ff., Kintsch 1974:45-70, Schnotz 1994:150-158, 163-168). Eine Proposition gilt hierbei als Wissenseinheit, die üblicherweise durch eine Relation und ihre Argumente dargestellt wird (vgl. Anderson 1995a:141 f.). In Anlehnung an Engelkamp (1984a:35 f.) und Estes (1994:5) wird der Begriff „Proposition" in dieser Arbeit als eine Art von Konzept betrachtet. Weitere Eigenschaften von Konzepten bzw. konzeptuellen Strukturen werden in Abschnitt 1.3.1 behandelt, indem untersucht werden soll, wie sie sich entwickeln. Diese Frage hängt eng mit der Funktionsweise der üblicherweise „Schema" genannten mentalen Vorrichtung zusammen, die den Rahmen für konzeptuelle Entwicklung bzw. Wissenserwerb bietet. In diesem Abschnitt wird auch gezeigt, dass die Reduktion der für den weiteren Prozess aufgenommenen Informationen nicht nur für die Indexierung wichtig ist, sondern auch ein entscheidendes Merkmal bei der Konzeptentwicklung darstellt.[8] Auf das Modell der Schemata ist auch die Annahme zurückzuführen, dass die konzeptuelle Struktur netzwerkartig strukturiert ist (vgl. z. B. Anderson

[8] Der reduktive Charakter der Wahrnehmung wird schon in der Gestaltpsychologie betont, deren Grundgedanke die Betrachtung der einzelnen Phänomene als „überpunktuelle Gebilde oder Sachverhalte, die räumlich, zeitlich oder raumzeitlich ausgedehnt sind, mit den Eigenschaften, die sich nicht aus artgleichen Eigenschaften der punktuellen Elemente herleiten lassen[,]" ist (Metzger 1954:125).

1995a:150, Klix 1988, Sowa 1984:76, Engelkamp 1985:293-296). Diese Annahme ist für diese Arbeit sehr relevant, weil sie die Basis für die systematische Untersuchung des Beitrags darstellt, den die durch die lexikalischen Relationen reflektierten konzeptuellen Relationen zum Aufbau der konzeptuellen Struktur von Texten leisten. Wissen und Information sind die Gegenstände, auf die sich Informationsverarbeitung bezieht (vgl. Rickheit/Strohner 1992:14 f.). Beide Begriffe werden in Abschnitt 1.3.2 behandelt. Was sich durch die Schemafunktion nicht erklären lässt, ist der genaue Zusammenhang zwischen Bedeutungen und Konzepten. In 1.3.3 werden einige Positionen zu diesem Problem behandelt.

Für die Untersuchung der durch lexikalische Relationen repräsentierten konzeptuellen Struktur von Texten stellen sich die folgenden Fragen, die in Kapitel 2 behandelt werden: Erstens, aufgrund welcher Eigenschaften können lexikalische Relationen für die Analyse von Texten unabhängig von ihren unterschiedlichen Inhalten eingesetzt werden? Zweitens, welche kognitiven Funktionen und Rollen stellen lexikalische Relationen im Text dar? Die Abschnitte 2.1 und 2.2 beziehen sich auf die erste, Abschnitt 2.3 auf die zweite Frage. In 2.1 werden allgemeine Eigenschaften der einzelnen lexikalischen Relationen vorgestellt. In der künstlichen Intelligenz (KI) ist es üblich, für die simulierten kognitiven Prozesse eine Wissensbasis zu verwenden, in der das dafür benötigte Wissen repräsentiert ist (vgl. Gašević/Djurić/Devedžić 2006:4-7). Man kann die anzusetzende Wissensbasis als eine Repräsentation der konzeptuellen Struktur des kognitiven Subjekts bzw. des menschlichen Indexierers und die zu analysierenden Texte als Träger der aufzunehmenden Informationen betrachten. Wissensrepräsentation kann also einerseits als Surrogat der menschlichen Wissensstruktur angesehen werden, andererseits aber auch als reales Medium für die darauf basierenden kognitiven Prozesse. Wissenschaftler sind sich nicht darüber einig, welche Arten von Wissen das Lexikon umfassen soll (vgl. Bußmann 2002:428). Für diese Arbeit relevant ist jedoch nur das Wissen, das im Lexikon durch lexikalische Relationen repräsentiert ist. In 2.2.1 wird gezeigt, dass Wortkonzepte auf der mentalen Ebene auch netzwerkartig strukturiert sind und lexikalische Relationen einen Teil der konzeptuellen Strukturen darstellen, die Menschen zur Verfügung stehen. Dies führt zu der weiteren für diese Ar-

beit relevanten Annahme, dass lexikalische Relationen ein sehr reduziertes Wissen darstellen. Das durch sie repräsentierte Bedeutungswissen ist im Vergleich zum episodischen oder sogar enzyklopädischen Wissen konstant und kontextunabhängig. Wenn dies zutrifft, dann können die lexikalischen Relationen als Wissensrepräsentation für die Analyse der Texte kontextunabhängig eingesetzt werden. Die Wissensbasis, die für diese Arbeit verwendet wird, ist das in Abschnitt 2.2.3 zu behandelnde WordNet (Version 2.0), das ein maschinenzugängliches Lexikon darstellt, in dem das Wissen über Wörter durch lexikalische Relationen zu anderen Wörtern repräsentiert ist.

In Abschnitt 2.3 geht es um die Rolle und Funktion der lexikalischen Relationen im Text. Die Rolle der lexikalischen Relationen wird in der Linguistik oft unter „lexikalischer Kohäsion“ abgehandelt. Neben Kohäsion ist „Kohärenz“ ein weiterer Begriff, der den Sinnzusammenhang von Texten bezeichnet. Beide Begriffe werden in der Literatur sehr uneinheitlich verwendet, wodurch oftmals erhebliche terminologische Verwirrung entsteht. In Abschnitt 2.3.1 werden einige Definitionsversuche vorgestellt, und im Anschluss wird der Begriff „Kohäsion“ präzisiert, damit er für die automatische Textanalyse nutzbar gemacht werden kann. Es gibt verschiedene Ansichten darüber, welche Arten von lexikalischen Relationen bei der Analyse von Texten berücksichtigt werden sollen. In Abschnitt 2.3.2 werden einige dieser Ansichten vorgestellt und diejenigen lexikalischen Relationen bestimmt, die in dieser Arbeit in Betracht gezogen werden. In den Abschnitten 2.3.3 und 2.3.4 wird untersucht, wie die lexikalische Kohäsion zur Erstellung und zum Aufbau der konzeptuellen Struktur im Text beiträgt: Wenn Kognition als Aktivität des Wissens aufgefasst wird (vgl. Neisser 1976:1), lassen sich in einer sprachlichen Kommunikation zwei Arten dieser Wissensaktivität unterscheiden: Textproduktion und Textverstehen. Der Abschnitt 2.3.3 bezieht sich auf die Textproduktion und untersucht, inwiefern die lexikalische Kohäsion einen Teil der Wissensstruktur bzw. konzeptuellen Struktur von Textproduzenten darstellt. Anschließend wird in Abschnitt 2.3.4 gezeigt, dass lexikalische Relationen wie Hyperonymie/Hyponymie und Meronymie/Holonymie einen Teil der konzeptuellen Struktur von Texten darstellen, durch die Informationsreduktion beim Textverstehen reflektiert wird.

In Kapitel 3 wird versucht, die im theoretischen Teil untersuchte kognitive Charakteristik der einzelnen lexikalischen Relationen auf die Indexierung anzuwenden. In Abschnitt 3.1 werden die drei Prinzipien der Indexierung behandelt, in Abschnitt 3.2 werden einige Komponenten des IRS vorgestellt. Die Implementierung der Anwendung der kognitiven Eigenschaften der lexikalischen Relationen auf die Indexierung und die dabei entstehenden Probleme werden in 3.3 dargestellt. Eines der kritischen Probleme, die sich auf die Eigenschaften der Daten (also der Wörter) zurückführen lassen, besteht darin, dass sich mehrere mögliche Konzepte einer Wortform zuweisen lassen (Polysemie). Daraus können wiederum mehrere mögliche Kohäsionsketten entstehen, denen ein Wort angehört. Ein weiteres Problem ist die Frage, wie die kognitiven Eigenschaften der lexikalischen Kohäsion bei der Repräsentation von Indextermen berücksichtigt werden können. Diesem Problem wird in dieser Arbeit durch die Bestimmung des semantischen Gewichts der Konzepte, das als Indextermgewicht verwendet wird, Rechnung getragen. Das nach der zuvor dargelegten Konzeption implementierte Indexsystem wird in Abschnitt 3.4 mit dem vektorraummodellbasierten Indexsystem verglichen, die mit den jeweiligen Ansätzen verbundenen Vor- und Nachteile werden untersucht.

1.3 Theoretische Vorbemerkungen

1.3.1 Konzept und Schema

Allgemeine Merkmale von Schemata

In der kognitiven Psychologie ist es eine gängige Annahme, dass sich konzeptuelle Entwicklung und Wissenserwerb am besten durch die „Schema" (bzw. „Frame") genannte mentale Vorrichtung erklären lassen (vgl. Mandl/Friedrich/Hron 1988:124-135, Rumelhart 1980, Schnotz 1994:61, Sowa 1984:42-51). Für unterschiedliche kognitive Bereiche wurden spezifische Schemamodelle entwickelt, die sich aber alle in der grundlegenden Funktionsweise gleichen, wie z. B. visuelle Perzeption (vgl. Neisser 1976), menschliche Handlung (vgl. Schank/Abelson 1997, Zimbardo 1988:73 ff.) oder allgemeine kognitive Entwicklung (vgl. Kamppinen 1993a:143-163). Schließlich werden Schemata auch verwendet, um die Funktionsweise des Gedächtnisses zu erklären (vgl. Schachter 1989:691-694). Neisser (1976:54) stellt die Funktionsweise von Schemata und ihrer Beteiligung an der Wahrnehmung wie folgt dar:

> „A schema is that portion of the entire perceptual cycle which is internal to the perceiver, modifiable by experience, and somehow specific to what is being perceived. The schema accepts information as it becomes available at sensory surfaces and is changed by that information; it directs movements and exploratory activities that makes more information available, by which it is further modified."

Eines der wichtigsten Charakteristika von Schemata ist, dass sie ständiger Veränderung unterliegen (vgl. Neisser 1976:20 ff., 56). Bei der aktuellen Wahrnehmung werden anhand der eingehenden Informationen und der vorhandenen schematischen Struktur Hypothesen über die Natur und die Eigenschaften des Wahrnehmungsgegenstandes gebildet. Diese führen zur gezielten und kontrollierten Aufnahme neuer Informationen, welche die Hypothesen mit der Zeit bestätigen, verbessern oder zu ihrer Verwerfung und der Aufstellung neuer Hypothesen führen. Ein Schema ist eine mentale Vorrichtung, die die Aufnahme von Informationen steuert, aber gleichzeitig durch diese Informationsaufnahme verändert wird. Dies macht den selbstreferenziellen Charakter der Schemata aus, der in Abbildung 1-1 veranschaulicht wird.

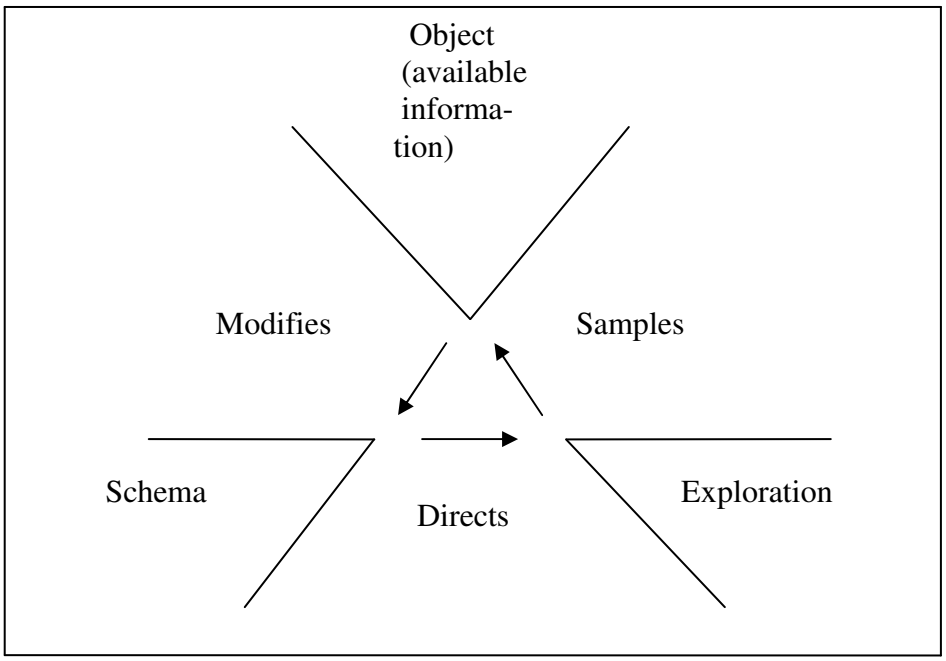


Abbildung 1-1: Schemazirkel (Neisser 1976:21)

Wie die Abbildung zeigt, leitet das Schema den Wahrnehmungsprozess („directs exploration“), der die am Wahrnehmungsobjekt verfügbaren Informationen auswählt („samples available information“), die wiederum das Schema modifizieren („modifies schema“). Es ist zu erkennen, dass die schematische Struktur nicht bloß ein leerer Rahmen für automatisch ablaufende Prozesse ist, sondern ein Rahmen, der fest verbunden ist mit seinen Inhalten, den Konzepten bzw. Propositionen (vgl. Engelkamp 1984a:36, Sowa 1984:42). Da Konzepte das Resultat der Informationsverarbeitung darstellen, können sie als Wissenselemente gelten (vgl. Abschnitt 1.2), und die Struktur der Schemata kann als „Wissensstruktur“ betrachtet werden (vgl. Rumelhart 1980:33). Der duale Charakter der Schemata, der darin besteht, gleichzeitig Repräsentation und Prozess zu sein, spielt nicht nur für die Gedächtnisforschung,[9] sondern auch für die KI eine entscheidende Rolle (vgl. Bobrow/Norman 1975).

Der Repräsentationsaspekt der Schemata

[9] Das Gedächtnismodell von Anderson (1983) besteht z. B. aus den drei Gedächtnissystemen deklaratives, Arbeits- und Produktionsgedächtnis und ihren Interaktionen. Hierher gehört außerdem die Differenzierung zwischen deklarativem und prozeduralem Wissen (vgl. Konerding 1993:84, Schnotz 1994:36 f.).

Hinsichtlich der Wissensrepräsentation wird die schematische Struktur oft als „Format“ oder als „Datenstruktur“ charakterisiert (vgl. Minsky 1975:212, Neisser 1976:55 f.). Jeder Mensch hat andere Erfahrungen mit einem bestimmten Sachverhalt gemacht, weshalb das Wissen über diesen Sachverhalt von Individuum zu Individuum unterschiedlich ist. Damit Individuen weiteres Wissen zum Beispiel durch sprachliche Kommunikation erwerben können, muss das individuelle Wissen der Gesprächsteilnehmer eine Schnittmenge besitzen. Hierzu gehören besonders die Stereotypen, die man als konventionalisiertes Wissen über Gattungsbegriffe bezeichnen kann (vgl. Konerding 1993:55-58). Dieses stereotypische Wissen drückt sich in sog. Default-Werten aus, die von Anderson (1995a:151) als „typische Ausprägungen auf den einzelnen Attributen“ eines Konzepts bezeichnet werden. Diese Default-Werte bilden die Ausgangshypothese für die Verarbeitung der neuen Information (vgl. Minsky 1975:212). Aus diesem Grund greifen zahlreiche Modelle der Wissensrepräsentation, die auf der Schematheorie basieren, auch auf den Stereotypenbegriff zurück (wie z. B. Anderson 1995a:150 ff., Dietze 1994:70-74, Konerding 1993:144-160). Bei Stereotypen ist die Informationsreduktion, die bei der Konzeptentwicklung stattfindet, darin zu erkennen, dass nur typische Merkmale den Inhalt des Konzepts bilden, während untypische Merkmale wie „dreibeinig“ oder „blau“ nicht in die entsprechende Konzeptstruktur von TIGER[10] bzw. ROSE aufgenommen werden. Die Struktur von Schemata („Frames“) hat Minsky (1975:212) weiter spezifiziert:

> „We can think of a frame as a network of node and relations. The ‘top levels’ of a frame are fixed, and represent things that are always true about the supposed situation.“

Dem Zitat ist die für diese Arbeit entscheidende Annahme zu entnehmen, dass ein Schema bzw. ein Frame eine netzwerkartige und hierarchische Struktur aufweist. Warum sind Schemata so strukturiert? Eine Antwort darauf kann man finden, wenn man den prozeduralen Aspekt der Schemata untersucht.

Der prozedurale Aspekt der Schemata

[10] Zur Schreibweise s. 1.3.4

„Assimilation“ und „Konstruktion“ sind zwei Prozesse, die immer für die Erklärung der Konzeptentwicklung bzw. des Wissenserwerbs herangezogen werden (vgl. z. B. Aebli 1988, Rumelhart/Norman 1978, Seel 2003:149-155, Seiler/Wannenmacher 1983a:320 f., Zimbardo 1988:72 ff.). Wenn die Eigenschaften des Wahrnehmungsobjektes mit der Merkmalsstruktur eines vorhandenen Konzepts einigermaßen oder ganz übereinstimmen, dann wird es erkannt.[11] Hier kommen die Relationen „Kontrast“ und „Ähnlichkeit“ ins Spiel, die zusammen mit der Kontiguität seit Aristoteles als grundlegende Assoziationstypen verstanden werden, durch die die einzelne Phänomene oder Ideen zu größeren und komplexeren gedanklichen Einheiten verknüpft werden (vgl. Blank 2001:38-44, Chaffin/Herrmann 1988:20, Raible 1981:1-7). Kontrast und Ähnlichkeit stehen zueinander in komplementärer Beziehung, da Ähnlichkeit bei minimalem Kontrast deutlicher wird und Kontrast zur Abhebung des Ähnlichen von Nichtähnlichem beiträgt. Kontiguität weist auf die Relation zwischen zusammen auftretenden Erlebniseinheiten hin und umfasst zum Beispiel physikalische Nachbarschaft, zeitliche Folge, Ursache und Wirkung sowie Teil und Ganzes. Ein Wahrnehmungssegment ist durch Kontrast von anderen Segmenten hervorgehoben und wird durch die Ähnlichkeit zu einem bereits erworbenen Konzept als eben dieses Konzept oder als ein Teil davon erkannt (vgl. Konerding 1993:94-98). Hierbei handelt es sich um Assimilation, da keine Änderung der vorhandenen konzeptuellen Struktur erfolgt. Beispiele für Assimilation sind Muster- und Objekterkennung. Es zeigt sich, dass Ähnlichkeit und Kontrast in der Sprache ihre Entsprechung in den lexikalischen Relationen Synonymie und Antonymie finden, auf die in Abschnitt 2.1 genauer eingegangen wird. Wenn die Assimilation nicht gelingt, also keine Ähnlichkeit zwischen den analysierten Informationen und den vorhandenen Konzepten besteht, dann erfolgt eine strukturelle Änderung, die man als „Konstruktion“ bezeichnet.[12] Als ein typisches Beispiel für die Konstruktion ist die Schemainduktion zu nennen (vgl. Rumelhart/Norman 1978:45 ff.): Hierbei wird aus gleichzeitig ak-

[11] Hier sind auch zwei in der kognitiven Semantik wichtige Begriffe – „Token“ und „Type“ – zu erwähnen, wobei Ersterer „einzelne Vorkommensfälle des Wahrnehmungssegments“ und Letzterer die „Gesamtheit von Segmenten, die (intraindividuell) über Ähnlichkeitsbeziehungen assoziativ verfügbar sind“, bezeichnet (vgl. Konerding 1993:120).

[12] Der Veränderungsgrad der schematischen Struktur ist von der Diskrepanz zwischen den neuen Informationen und der vorhandenen schematischen Struktur abhängig. Dies entspricht dem Äquilibrationsprinzip von Piaget, das besagt, dass sich die geistige Entwicklung durch das Suchen und Finden von mentalen Gleichgewichtszuständen vollzieht (vgl. Montada 2002:438 f.).

tivierten Schemata, die zum Beispiel auf bestimmte zeitliche, räumliche oder logische Relationen hinweisen, ein neues Schema generiert, das die aktivierten Schemata als Subschemata subsumiert (vgl. Rumelhart/Norman 1978:45 ff.). An diesem Beispiel zeigt sich, dass durch das Inbeziehungsetzen verschiedener Schemata eine Netzwerkstruktur entsteht. [13]

Die Frage, warum sich die Netzwerkstruktur besonders durch einen hierarchischen Charakter auszeichnet, kann vorläufig durch das folgende Zitat von Minsky (1975:212) beantwortet werden, der auf eine wichtige Wirkung bei der Transformation der Schemata hinweist:

> „The effects of important actions are mirrored by transformations between the frames of a system. These are used to make certain kinds of calculations economical, to represent changes of emphasis and attention, and to account for the effectiveness of imagery".

Also sind nicht alle Merkmale der jeweiligen Subschemata an der Konstruktion der neuen Schemata beteiligt, sondern nur die relevanten Merkmale. Diese generalisierende Funktion von Schemata führt zu „Konzepten" als den „in unserem Gedächtnis gespeicherte[n] Einheiten, die Information über Abschnitte unserer Umwelt zusammenfassend widerspiegeln" (Hoffmann 1986:56). Im Abschnitt 2.3.4.2 wird näher untersucht, wie sich der reduktive Charakter der Konzeptentwicklung durch hierarchische Relationen wie IS-A und HAS-A auf der Repräsentationsebene widerspiegelt.

1.3.2 Wissen, Information

In Abschnitt 1.3.1 wurde davon ausgegangen, dass ein Reiz der Außenwelt zu Information wird, wenn er selektiv in den Wahrnehmungsprozess aufgenommen wird, und dass die verarbeitete Information zu Wissen wird, wenn sie den anschließenden kognitiven Prozessen bzw. Informationsprozessen als Basis

[13] Bühler (1966:60-73) hat drei Grundtypen von Bewusstseinsweisen bzw. Gedankenformen unterschieden, nämlich „Regelbewusstsein", „Beziehungsbewusstsein" und „Intentionen". Der zweite Typ „Beziehungsbewusstsein" behält seinen Stellenwert auch in der Bestimmung des Denkens bei Aebli (1988:228 ff., 2001:19) bei, aber mit Betonungsverlagerung: Aebli behauptet, dass Denken ein Vorgang ist, bei dem eine Beziehung zwischen Denkenselementen hergestellt wird. Nach dieser Ansicht deckt der Begriff „Denken" nicht nur Urteilen, Erinnern und Schätzen ab, die normalerweise bewusst geschehen, sondern auch Objekterkennung, die im Alltag oft unbewusst geschieht.

dient. Von Wissen kann erst als Resultat der informationsverarbeitenden Prozesse die Rede sein, während Informationen die Entitäten sind, die dem Wahrnehmungsprozess zugrunde liegen, solange er noch in Gang ist.[14] Schwierig wird es, wenn man versucht, Wissen und Information auf der kognitiven Ebene nur unter dem Repräsentationsaspekt zu unterscheiden.[15] Dasselbe Problem tritt auf, wenn die Begriffe „Wissen" und/oder „Information" auf Medien[16] übertragen werden, die selbst keine Verarbeitungsfähigkeit besitzen. Eine Differenzierung kann erfolgen, wenn man mit Buckland (1991:50) eine Abhängigkeit der Eigenschaft der physikalischen Information von der Situation sieht: Die typische Situation, in der Wissen bzw. Information durch physikalische Medien selektiert und aufgenommen wird, ist Kommunikation, als deren spezieller Fall Information Retrieval (IR) angesehen werden kann.[17] Für Kommunikationssituationen, die durch die Übermittlung von Texten gekennzeichnet sind, können weitere Überlegungen über Wissen und Informationen angestellt werden: Damit jemand sein Wissen dem anderen Kommunikationsteilnehmer mitteilen kann, produziert er einen Text. Dies ist der Moment, in dem das Wissen auf ein Medium übertragen wird. Das auf das Medium übertragene Wissen wird zu Information, wenn der Kommunikationspartner, also der Leser oder Hörer, das Wissen des Textproduzenten aufnimmt und verarbeitet. Wenn diese Information vom Textleser verarbeitet und gespeichert wird, kann sie wieder als Wissen angesehen werden. Also kann dieselbe Entität je nach Perspektive und Rolle der Kommunikationsteilnehmer als Wissen oder Information be-

[14] Hier exemplarisch weitere Definitionen von Information und Wissen: „Information is the designation of the content obtained from the external world in the process of our adaptation to it and the adaptation of our sense to it." (Frants/Shapiro/Voiskunskii 1993:43, zitierend Wiener 1954); „information is data that changes the state of a system that perceives it." (Meadow/Boyce/Kraft 2000:37); „Wissen lässt sich […] definieren als das, was einem Agenten zuzuschreiben ist, damit sein Verhalten […] berechnet und erklärt werden kann." (Konerding 1993:83, zitierend Newell 1982)

[15] Dieses Problem wird durch das folgende Zitat von Tergan (1989a:155) deutlicher: „Mit ‚Wissen' werden […] alle in irgendeiner Weise mental repräsentierten Informationen bezeichnet."

[16] Hier sind „Medien" im Sinne von physikalischem Material gemeint, in dem das Wissen enkodiert ist.

[17] Buckland (1991:94 f.) hat den charakteristischen Aspekt, durch den sich IR von konventioneller Kommunikation abhebt, wie folgt beschrieben: „Within the communication, retrieval-based information services can be seen as constituting a special class in which the fate of the message is lost to the communicators and is controlled successively by the provider of the retrieval-based system, who decides which message will be stored and how they will be retrieved and the users, who determine which of the retrievable messages will be received and which will be read."

trachtet werden. Aber ohne Bezug auf eine konkrete Situation handelt es sich bei Information und Wissen um dieselbe Entität.

1.3.3 Bedeutung und Konzept

Wenn eine lexikalische Relation die Relation ist, die zwischen zwei Wörtern besteht, dann stellt sich die Frage, auf welchen Bedeutungsaspekt oder auf welche Bedeutungsdimension von Wörtern sie sich bezieht. Lyons unterscheidet zwischen „deskriptiver" und „nicht-deskriptiver Bedeutung", wobei Letztere weiter in „soziale" und „expressive Bedeutung" unterteilt werden kann (vgl. Lyons 1995:44 f., Lyons 1977:64-70). Von „sozialer Bedeutung" ist die Rede, wenn sprachliche Ausdrücke auf soziale Beziehungen oder Handlungen verweisen; die „expressive Bedeutung" bezieht sich auf subjektive Gefühle, Bewertungen etc. „Deskriptive Bedeutung" bezieht sich auf den inhaltlichen Aspekt der Bedeutung, anhand dessen sprachliche Ausdrücke auf einen Sachverhalt referieren können, der als wahr oder falsch beurteilt werden kann. Weiter bestimmt Lyons (1977:281ff.) denjenigen Bedeutungsaspekt als „Sinn" (engl. „sense"), auf dem lexikalische Relationen beruhen. Während „Sinn" oft synonym zu „deskriptiver Bedeutung" verwendet wird (vgl. Cruse 2000:22, Leech 1981:26), versteht Lyons (1977:281-327) darunter nur den Aspekt der deskriptiven Bedeutung von Ausdrücken, der die „Sinnrelationen" zwischen einem Ausdruck einer Einzelsprache und anderen Ausdrücken dieser Sprache umfasst.

Die kognitive Semantik untersucht Bedeutung auf der mentalen Ebene und greift hierfür oft auf den Begriff des Konzepts zurück (vgl. z. B. Engelkamp 1983, Seiler/Wannenmacher 1983a, Schwarz 2002). Ein Konzept umfasst im Prinzip alles, was von einem Individuum über einen Gegenstand bzw. Sachverhalt gewusst werden kann. Aufgrund dieses dynamischen Charakters erweisen sich Konzepte als prinzipiell nicht abgeschlossene Entitäten (vgl. Abschnitt 1.3.1). Der Terminus „Konzept" ermöglicht daher auch die Berücksichtigung von kontextuellen Bedeutungsfaktoren, die von der Sprache unabhängig sind.
Wie Konzept und Wortbedeutung auf der mentalen Ebene aber konkret repräsentiert sind und miteinander interagieren, ist in der kognitiven Semantik eine

äußerst umstrittene Frage. Bei der Frage nach der Relation zwischen Bedeutung und Konzepten auf der mentalen Ebene lassen sich üblicherweise zwei verschiedene Positionen unterscheiden (vgl. Schwarz 2002, Levinson 1997): Die Wissenschaftler, die einen holistischen Ansatz vertreten, sind der Meinung, dass sich sprachbezogenes semantisches Wissen nicht von Weltwissen unterscheidet, und setzen daher Wortbedeutung und Konzepte auf der kognitiven Ebene gleich (vgl. z. B. Cruse 2000:127 ff., 2002a:543 ff., Engelkamp 1985:292, Jackendoff 1983:95, Langacker 1987:5, Löbner 2003:23 ff.). Dennoch heißt dies nicht unbedingt, dass Konzepte alle Dimensionen der Bedeutung abdecken. Zum Beispiel unterscheidet Leech in Hinblick auf den kommunikativen Wert (engl. „communicative value") zwischen konzeptueller, assoziativer und thematischer Bedeutung, wobei er die konzeptuelle Bedeutung als den wichtigsten Aspekt der sprachlichen Kommunikation betrachtet (vgl. Leech 1981:10-27). Die konzeptuelle Bedeutung wird von Cruse direkt mit der deskriptiven Bedeutung (s. o.) in Verbindung gebracht,[18] insofern der konzeptuelle Teil der Wortbedeutung dadurch entsteht, dass ein einzelnes Konzept einer Wortform zugewiesen wird (vgl. Cruse 2002a:543 ff.). Für Löbner decken Konzepte explizit nur den deskriptiven Aspekt der Bedeutung ab, und er betrachtet lexikalische Relationen als „Bedeutungsrelationen" (vgl. Löbner 2002:29, 116-135).

Wie oben erwähnt, sind Konzepte offene Entitäten, die je nach dem Erfahrungshintergrund der kognitiven Subjekte unterschiedlich sein können. Andererseits ist die Existenz relativ konstanter Teile von Konzepten, die den Mitgliedern einer Sprachgemeinschaft in gleicher Form zur Verfügung stehen, unumgänglich für das Funktionieren von Kommunikation und damit für eine wichtige Art des Wissenserwerbs.[19] Aus diesem Grund vertreten einige Wissenschaftler, darunter vorwiegend Linguisten, die Modularitätsannahme, der

[18] Über den Zusammenhang zwischen konzeptueller und deskriptiver Bedeutung schreibt Cruse (2000:47) folglich: „[descriptive meaning] is fully conceptualized. That is to say, it provides a set of conceptual categories into which aspects of experience may be sorted. Such a categorization effectively describes the experiences and licenses further inferences about their properties, and so on."

[19] Langacker (1987:159) erklärt die Existenz eines relativ konstanten Teils von Konzepten durch die sog. „Zentralität" (engl. „centrality"), die er wie folgt beschreibt: „I do not specifically claim that all facets of our knowledge of an entity have equal status, […]. The multitude of specification that figure in our encyclopedic conception of an entity clearly form a gradation in terms of their centrality. Some are so central that they can hardly be omitted from even the sketchiest characterization, whereas other are so peripheral that they hold little significance even fort he most exhaustive description."

zufolge auf der kognitiven Ebene eine semantische Komponente getrennt von den konzeptuellen Strukturen existiert, in der die kontextunabhängigen und konstanten sprachlichen Informationen gespeichert sind (vgl. Bierwisch 1983:75-94, Levinson 1997:14, Schwarz 2002:280 ff.). Nach dieser Ansicht von der Autonomie des Sprachsystems wird streng unterschieden zwischen Weltwissen und sprachlichem Wissen, das von Sprache zu Sprache unterschiedlich ist. Die Auffassung von Aitchison (1994:52 ff.) kann zwischen den beiden Positionen eingeordnet werden: Sie glaubt, dass die Wortbedeutung zwar einen großen Teil, aber nicht das ganze Konzept abdeckt; das Konzept umfasst auch Bedeutungsbereiche, die normalerweise nicht vom Wort erfasst werden. Nach dieser Auffassung bezieht sich Wortbedeutung auf den relativ stabilen Teil von Konzepten, die den Mitgliedern einer Sprachgemeinschaft gemeinsam zur Verfügung stehen. Die Annahme, dass kontextunabhängige konstante Teile der Konzepte existieren, rechtfertigt in der vorliegenden Arbeit die Verwendung der lexikalischen Relationen als kontextneutrale Wissensbasis bei der Textanalyse (vgl. Abschnitt 2.2.2), da lexikalische Relationen üblicherweise von jedem Mitglieder einer Sprachgemeinschaft kontextunabhängig erkannt werden. Für die Computerlinguistik ist die Modularitätsannahme brauchbarer, da sie ein theoretisches Modell für die Wissensklassifikation anbietet, das einen wohlstrukturierten Systementwurf ermöglicht.

1.3.4 Wort, Lexem und lexikalische Einheiten

Wörter gelten als Grundeinheiten der Kommunikation, die stabil in unserem Wortschatz gespeichert sind (vgl. Schippan 2002:1). Dennoch ist es nicht ganz einfach zu bestimmen, was ein Wort ist, da verschiedene sprachliche Informationen wie morphologische, syntaktische, lexikalisch-semantische und phonologische Merkmale dabei als zu berücksichtigende Faktoren ins Spiel kommen (vgl. z. B. Bußmann 2002:750, Lewandowski 1994:1247 ff., Lipka 2002:88 ff.). In einer ersten Annäherung können die im Text vorkommenden Einheiten, die graphisch durch Leerstellen von anderen Einheiten abgetrennt sind, als Wörter verstanden werden. In dieser Arbeit ist die Definition von Wörtern als „d[en] kleinsten relativ selbständigen Träger[n] von Bedeutung, die im Lexikon kodifiziert sind“ (Bußmann 2002:750), von besonderem Be-

lang, wobei von morphosyntaktischen Merkmalen wie Deklination und Konjugation abstrahiert wird. Auf die Frage, was eigentlich diese kleinsten relativ selbstständigen Bedeutungsträger sind, sind zwei Antworten vorstellbar (vgl. Schindler 2002:33-44). Man kann bei dieser Definition eine gewisse Übereinstimmung finden mit den Eigenschaften entweder einer lexikalischen Einheit als Einheit von Form und Bedeutung oder auch eines Lexems, das als die Menge der lexikalischen Einheiten zu bestimmen ist. Zum Beispiel sind für die Wortform „Saturn“ in WordNet folgende zwei Bedeutungen eingetragen:

1. (3) Saturn – (a giant planet which is surrounded by three planar concentric rings of ice particles; 6th planet from the sun)
2. Saturn (Roman mythology) god of agriculture and vegetation; counterpart of Greek Cronus; „Saturday is Saturn’s Day“)

Das Lexem „Saturn“ umfasst außer der Wortform auch beide Bedeutungen, die das Potenzial besitzen, der Wortform zugewiesen zu werden. Von daher können Polyseme unter einem Lexem berücksichtigt werden. „Lexikalische Einheit“ ist die Kombination von einer Wortform und einer möglichen Bedeutung. Im obigen Fall sind zwei lexikalische Einheiten gegeben. Da eine lexikalische Relation auf nur einer möglichen Bedeutung von Wörtern beruht, bezieht sich der Begriff „Wort“ bei der Betrachtung der lexikalischen Relationen auf die lexikalische Einheit.

Da Lexeme und lexikalische Einheiten in der Linguistik üblicherweise als abstrakte Grundformen („types“) für verschiedene morphosyntaktische Realisierungsformen („tokens“) verwendet werden, gehören die Wörter, die zu einem Lexem oder einer lexikalischen Einheit gehören, zur gleichen Wortklasse. Semantische Ähnlichkeit ist aber auch unter Wörtern zu finden, die zu unterschiedlichen Wortklassen gehören. Zum Beispiel sind die drei Wörter *success*, *suceed* und *successful* auf irgendeine Weise semantisch ähnlich, aber sie gehören zu verschiedenen lexikalischen Kategorien. Solche Wortgruppen lassen sich besser mit Blick auf die Wortbildung erklären und werden oft als „Wortfamilie“ oder „Lexemverband“ aufgefasst. Wortfamilien bzw. Wortverbände können durch verschiedene Mittel wie Komposition oder Derivation entstehen (vgl. Lipka 2002:101, Schippan 2002:43). Damit werden Wörter bezeichnet, die gemeinsame Wurzelmorpheme besitzen

Der Begriff „Wort“ wird in dieser Arbeit im allgemeinen Sinne verwendet, solange seine Verwendung im jeweiligen Kontext unmissverständlich ist und kein Bedarf vorhanden ist, ihn explizit zu spezifizieren. Wenn von einem Wort in diesem Sinne die Rede ist, dann wird das durch kursive Schreibweise gekennzeichnet (z. B. *apple*). Wenn auf ein Wort im Sinne eines Lexems oder einer lexikalischen Einheit Bezug genommen werden soll, dann ist es im ersten Fall mit spitzen Klammern (<apple>) und im zweiten Fall mit einfachen Anführungszeichen (‚apple’) gekennzeichnet. Die Differenzierung von Form- und Bedeutungsebene, die Wortform und Konzept entsprechen, wird in dieser Arbeit streng eingehalten. Eine Wortform ist mit „“ markiert („apple“) und ein Konzept in Großbuchstaben geschrieben (APPLE).

Während Lexem und lexikalische Einheit in der Sprachwissenschaft relativ strikt definiert sind, scheinen die im Bereich des IR geläufigen Ausdrücke „Term“ und (lexikalisches) „Item“ mehr oder weniger fachübergreifende Termini zu sein. „Term“ bezieht sich manchmal auf die Wortform (vgl. Gaus 2003:56), kann aber auch die Bedeutungsebene mit einschließen (vgl. Bauer/Faschinger 2003). Es ist schwierig, solche linguistischen Begriffe mit dem Begriff Term, so wie er im Bereich des IR verwendet wird, in Einklang zu bringen. „Term“ wird in dieser Arbeit ohne weitere Berücksichtigung der linguistischen Eigenschaften als Synonym zu „Wort“ im allgemeinen Sinne verwendet. Zusammengefasst werden in dieser Arbeit also folgende Schreibweisen verwendet:

Wort (und Satz)[20] im allgemeinen Sinne (=Term): *apple*;
Wortform: „apple“;
Mit einer Wortform (und Satzform) verbundenes Konzept: APPLE;
Wort im Sinne von Lexem: <apple>;
Wort im Sinne von lexikalischer Einheit: ‚apple’.

[20] Die für Wörter im allgemeinen Sinn vorgesehene Notationskonvention gilt im Folgenden auch für Sätze, die Realisierungen von Propositionen darstellen (s. vor allem Abschnitt 2.3.4.2).

2. Kognitive Interpretation der lexikalischen Relationen

2.1 Lexikalische Relationen

Synonymie

Wesentlicher Charakterzug der Synonymie ist, dass sich die semantischen Eigenschaften der Wörter zum großen Teil überschneiden und dass der Kontrast zwischen Synonymen sehr schwach ist, wobei entweder keine oder eine nur sehr geringe Opposition zu finden ist (vgl. Cruse 1986:266, Lyons 1977:296). Für die Erkennung der Synonymie wird oft der sog. Austauschtest verwendet (vgl. Cruse 1986:268, 2002b:487-491). Zum Beispiel können *die*, *pass away, kick the bucket* und *pop one's clogs* als Synonyme erkannt werden, insofern sich die Bedeutung des Satzes *The old man died.* durch den Austausch von *died* durch *passed away*, *kicked the bucket* oder *popped his clogs* nur wenig verändert.[21] Der Austauschtest macht deutlich, dass viele Aspekte der Bedeutung bei der Untersuchung der Synonymie in Betracht gezogen werden müssen. *Die*, *pass away*, *kick the bucket* und *popped his clogs* stellen in deskriptiver Hinsicht denselben Sachverhalt dar, aber sie unterscheiden sich voneinander in expressiver Hinsicht. Es bestehen verschiedene Arten der Synonymie je nach dem Bedeutungsaspekt. Wenn die beiden Wörter in Hinsicht auf alle Aspekte der Bedeutung in einem Satz oder Kontext identisch sind, dann besteht absolute Synonymie (vgl. Cruse 1986:268, Lyons 1995:61). Hier überlappen sich die Merkmalsklassen der jeweiligen Wörter vollkommen. Die absolute Synonymie ist selten; viele Wissenschaftler behaupten sogar, es sei streng genommen keine vollkommene absolute Synonymie vorhanden (vgl. Cruse 1986:270, Lehrer 1974:23). Weiter werden von Cruse (1986:270-289) neben der absoluten Synonymie noch kognitive Synonymie und Quasisynonymie unterschieden. Kognitive Synonymie ist nach ihm als die Relation der Wörter definiert, die in ihrer deskriptiven Bedeutung gleich sind, aber in Hinsicht auf andere Aspekte wie expressive Bedeutung, stilistisches Register oder kollokationale Kombinierbarkeit unterschiedlich sind (vgl. Abschnitt 2.3.1). Kritisch sind die sog. Quasisynonyme (engl. „near-synonyms"), deren Austausch die Satzbedeutung, das heißt den Wahrheitswert des Satzes, nur in bestimmten Kontexen nicht verän-

[21] Problematisch ist bei dem Austauschtest der sog. logische Zirkel. Man muss nämlich erst wissen, ob die Wörter Synonyme sind, um überhaupt feststellen zu können, ob die Sätze semantisch identisch sind oder nicht (Cruse 1986:269).

dert, in anderen hingegen schon. Quasisynonyme sind z. B. *kill*:*murder*, *foggy*:*misty*, und *handsome*:*pretty.* Cruse (1986:287 ff., 2002b:493) hat versucht, dieses Problem anhand der Konfiguration von zentraler und peripherer Bedeutung in einem Kontext zu erklären. Der Satz *X is not murdered* impliziert, dass X nicht von jemandem getötet wurde, nicht jedoch, dass X nicht tot ist. Quasisynonymie beruht auf der peripheren Bedeutung der Wörter eines Wortpaars, sie liegt daher zwischen Synonymie und Nichtsynonymie. Daraus kann man weiter schließen, dass die zwischen Wörtern bestehende Ähnlichkeit unterschiedliche Granularität haben kann. Wie die Beispielkette *mound*/*hillock*/*hill*/*mountain* bei Cruse (1986:288) zeigt, steht jedes nebeneinander liegende Wortpaar unter Quasisynonymie, während dies bei den Wortpaaren, die nicht direkt benachbart sind, nicht der Fall ist. Von daher sind nur die Relationen der absoluten und der kognitiven Synonymie transitiv, die der Quasissynonymie aber nicht. Symmetrisch sind hingegen sowohl absolute und kognitive Synonymie als auch Quasisynonymie.

Hyperonymie/Hyponymie

Die gängige Beschreibungsmethode von Hyperonymie/Hyponymie ist Klasseninklusion (vgl. Cruse 1986:87, Lyons 1977:301). Während ein Hyperonym in extensionaler Hinsicht seine Hyponyme einschließt, ist das Inklusionsverhältnis in intensionaler Hinsicht umgekehrt. Zum Beispiel ist *elephant* Teil der Extension von *animal,* insofern *elephants* eine Untermenge von *animals* darstellen, während *animal* in der Intension von *elephant* enthalten ist, insofern *elephant* mehr semantische Merkmale als *animal* besitzt, es also semantisch impliziert (*elephant* $\supset$ *animal*). Logisch gesehen ist fast jede Synonymie Hyperonymie, weil die Synonyme einander einschließen können. Synonymie kann durch $A \neq B \ \& \ A \supset B \ \& \ A \subset B$ dargestellt werden, wobei solche bilateralen Relationen normalerweise Hyperonymie/Hyponymie ausschließen. In diesem Zusammenhang ist zu erkennen, dass Hyponymie nicht symmetrisch ist. Dass Hyponymie transitiven Charakter hat, verursacht ein kritisches Problem bei der automatischen Erkennung der lexikalischen Kohäsion, das in Kapitel 3 behandelt wird.

Die Inklusionsrelation führt zu natürlichsprachlichen Ausdrücken wie *An **X** is **Y*** oder *An **X** is a kind/type/sort of **Y*** (vgl. Cruse 1986:88 f., Lyons 1977:302). Aber dieser Paraphrasierungstest zur Überprüfung von Hyperonymie birgt eine Problematik, die anhand der folgenden Beispielssätze deutlich wird:

1) A sparrow is a kind of bird.
2) A dog is a kind of pet.
3) A dog is an animal.

Bei 1) gibt es den klaren Fall, dass die Zugehörigkeit kontextunabhängig klar ist, während 2) zeigt, dass Hunde im normalen Stadtleben zwar als Haustier akzeptiert sind, aber ohne Schwierigkeiten Ausnahmen zu finden sind. Dies führt zur Auffassung, nach der Hyponymie prototypischen Charakter besitzt (vgl. Cruse 2000:152). Cruse (1986:137-145) meint, dass mehr dominante und differenzierende Relationen als andere unter den Hyponymen zu bestimmen sind, die die sog. „Taxonomie" charakterisieren, einen aus der Botanik und der Zoologie übernommenen Begriff, der ein klassifikatorisches System beschreibt. Taxonomie gilt als besonderer Fall von Hyponymie, nämlich als prototypische Hyponymie. Da die in der taxonomischen Hierarchie befindlichen Wörter konstantere semantische Eigenschaften haben, brauchen sie normalerweise nicht mit adjektivischen Modifikatoren paraphrasiert werden (vgl. Cruse 1986:140). Dies gilt nicht nur für Wörter für natürliche Gattungen, sondern auch für Artefakte:

4) A violin is a musical instrument.
5) A seat is a furniture.

Dass ein Wort durch Klasseninklusion das Hyperonym von mehreren Wörtern sein kann, die untereinander im Verhältnis der Kohyponymie stehen, zeigen die Paare *apple*:*Baldwin* und *apple*:*Cortland*, aber auch zwei nicht unter Kohyponymie stehende Wörter können dasselbe Hyperonym haben, wie *novel*:*book* und *paperback*:*book*. *Book*:*paperback* und *book*:*novel* gehören verschiedenen Bedeutungsdimensionen (kognitiven Domänen) an, *book* ist hier also polysem.

Hyponymie kann nicht nur bei Nomen, sondern auch bei Adjektiven und Verben vorliegen. Da Verben und Adjektive einer anderen Wortart als Nomen angehören, kann die Überprüfungsmethode für Nomen („welche Art von") nicht

einfach auf diese übertragen werden. Bei diesen beiden Wortarten kann Hyperonymie durch die Frage „wie …“ oder „in welcher Weise …“ erkannt werden (vgl. Lyons 1977:304). Miller und Fellbaum führen zur Bezeichnung der Hyponymie bei Verben den Terminus „Troponymie“ ein und betrachten diese als eine Art der lexikalischen Implikation (vgl. Fellbaum 1998b:79 f., Miller/Fellbaum 1992:216-224). Der Begriff der Implikation, der eigentlich ein logisches Verhältnis zwischen Propositionen bezeichnet, kann auch zur Beschreibung von Relationen zwischen Verben benutzt werden, wenn zwei Sätze sich nur durch ihre Verben unterscheiden. Die daraus entstandenen Propositionen unterscheiden sich nur durch die Bedeutungen der Verben voneinander und können wegen deren Bedeutungen logische Implikationen bilden. Zum Beispiel impliziert der Satz *he is snoring* den Satz *he is sleeping*, und *he succeeded* impliziert *he tried*. Dadurch wird es auch möglich, die lexikalische Implikation von Verben zu berücksichtigen. Außerdem kann die lexikalische Implikation bei Verben nach dem Merkmal der temporalen Inklusion weiter subklassifiziert werden, wobei deren Grad für die Bestimmung der Troponymie eine Rolle spielt.

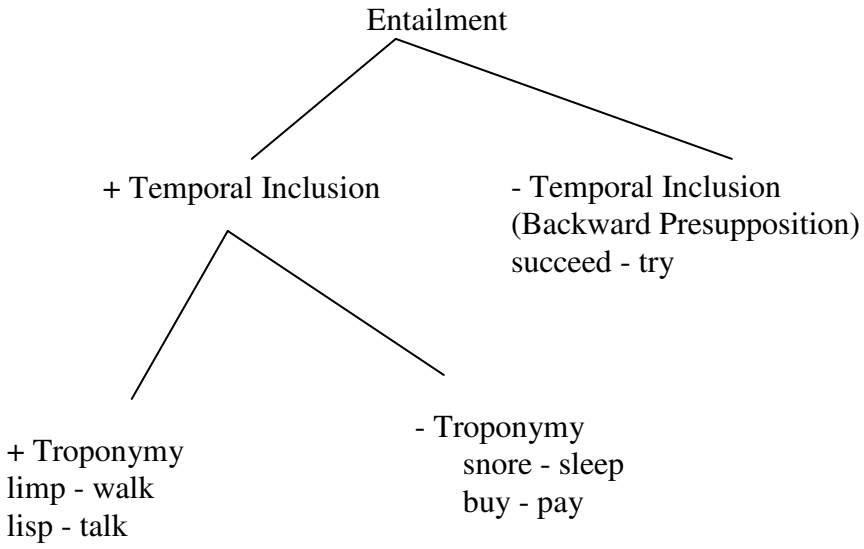


Abbildung 2-1: Drei Arten der Implikation von Verben (Miller/Fellbaum 1992:222)

Wie Abbildung 2-1 darstellt, besteht keine zeitliche Inklusion beim Paar *succeed/try*, da *succeed try* zeitlich nachfolgt. Die Fälle, in denen die von den je-

weiligen Verben denotierten Ereignisse sich zeitlich überschneiden, können nach dem Grad dieser Überschneidung weiter klassifiziert werden. Wenn das von einem Verb denotierte Ereignis mit dem von einem anderen Verb denotierten Ereignis zeitlich deckungsgleich ist und sich nur hinsichtlich der Quantität der Informationen über die Art und Weise unterscheidet, dann stehen die beiden Verben im Verhältnis der Troponymie, die als Hyperonymie zwischen Verben zu betrachten ist.

Opposition

Kontrast als eine der drei grundlegenden assoziativen Relationen neben Kontiguität und Ähnlichkeit (vgl. Abschnitt 1.3.1) wird in der Linguistik meistens unter dem Begriff „Opposition" behandelt (vgl. Cruse 1986:197, Lyons 1977:281-300). Der hauptsächliche Unterschied zwischen Kontrast und Opposition liegt allerdings darin, dass die prototypische Opposition binär ist, während Kontrast unabhängig von der Anzahl der kontrastierenden Elemente ist (vgl. Lyons 1977:289). Einige Wissenschaftler sind der Meinung, dass, wenn eine Äußerung wahrgenommen wird, ihr Gegenteil in gewisser Weise auch aktiviert ist (vgl. Lyons 1977:281). Cruse vermutet (2000:167), dass sich Opposition unmittelbar nach dem Lernen der Wörter ergibt und sie daher zu den kognitiven Primitiven zu zählen ist. An dieser Stelle sind noch die beiden Begriffe „Kontradiktion" und „Kontrarität" einzuführen, die beide Gegensatzbeziehungen bezeichnen. Während von Kontradiktion die Rede ist, wenn von zwei Propositionen eine nur dann wahr sein kann, wenn die andere falsch ist und vice versa, können bei der Kontrarität beide Propositionen wahr oder beide falsch sein (vgl. Lyons 1977:283). Diese Unterscheidung führt auf der Wortebene zur Unterscheidung von Komplementarität und Antonymie, wobei Erstere als prototypische Opposition gilt. Komplementäre Wortpaare wie beispielsweise *boy*:*girl*, *true*:*false*, *pass*:*fail* und *inside*:*outside* schließen einander logisch aus und schöpfen die Bedeutungsdimension aus, die Komplementarität ausmacht. Dieser ausschließende Charakter hängt mit der Abwesenheit der Graduierbarkeit in der Bedeutungsdimension zusammen, auf der die Komplementarität fußt. Eine Komparation wie *John ist deader than Tom* ist bei Wörtern eines Komplementärpaars in normalen Kontexten nicht akzeptabel, da der Komparativ die Eigenschaft der Gradierung voraussetzt. In diesem Sinne ist Komple-

mentarität streng binär, deswegen kann sie als prototypische Opposition angesehen werden.

Graduierbarkeit spielt auch eine wichtige Rolle für die Klassifikation der Antonymie (vgl. Cruse 1986:202 ff., Lyons 1977:283-289). Während Graduierbarkeit wegen der klaren Unterscheidbarkeit bei Komplementarität nicht vorkommt, kann es sie bei antonymen Paaren geben. Zum Beispiel ist es unmöglich, eine absolute Grenzlinie zwischen *long* und *short* oder *small* und *big* zu ziehen. Bei der Antonymie ist ein Bereich zwischen beiden Polen zu finden, dem sich beide Wörter nicht klar zuordnen lassen. Es kann etwa von einer Geschwindigkeit die Rede sein, die weder schnell noch langsam ist. Die Negation eines Wortes impliziert nicht unbedingt das Gegenwort. Deswegen ist Antonymie eine von Kontrarität geprägte Beziehung.

Weiter kann man die Graduierung mit der Komparativkonstruktion in Verbindung bringen und mit ihrer Hilfe verschiedene Antonymietypen unterscheiden (vgl. Cruse 1986:204-213). Zum Beispiel impliziert *A is heavier than B* nicht unbedingt, dass A schwer ist, während *A is colder than B* die Eigenschaft enthält, dass A kalt ist. Der Unterschied liegt darin, ob die semantischen Eigenschaften des Wortes auch im Komparativ erhalten bleiben oder nicht. In diesem Zusammenhang unterscheidet Cruse (1986:206 ff.) zwischen wahrem Komparativ und Pseudokomparativ, wobei im ersten Fall die eigentliche Bedeutung beim Komparativ verbleibt bzw. an ihn gebunden ist (engl. „committed“), während sie im anderen Fall neutral (engl. „impartial“) ist und bei der Komparation wegfallen kann. Polare Antonymie besteht zwischen Wortpaaren wie *long*:*short*, *heavy*:*light*, *fast*:*slow*, die nur Pseudokomparative bilden können. Die beiden Wortpaare sind in ihren komparativen Konstruktionen neutral, deshalb bedeutet *A is longer/shorter than B* nicht unbedingt, dass A einigermaßen lang bzw. kurz ist. Dies führt dazu, dass *long* in dem „Wie-Fragetest“ (*how long is A?*) die neutrale Dimension von LÄNGE darstellt; der Fragesatz enthält also keine Information darüber, ob A lang oder kurz ist. Die äquipollente Antonymie ist hingegen die Relation der Wortpaare, die wahre Komparative bilden können. Dazu gehören *nice*:*nasty*, *sweet*:*sour* und *happy*:*sad*. Beide Teile des jeweiligen Wortpaars können zweifellos für den „Wie-Fragetest“ verwendet werden. In diesem Fall implizieren die Wie-Fragesätze *How nice is he?* oder *How happy is he?*, dass die fragliche Person nett bzw. glücklich ist, weil

nice und *happy* anders als bei der polaren Antonymie nicht die neutrale Dimension von „Nettigkeit“ oder „Fröhlichkeit“ bezeichnen.
Überlappende Antonyme wie *good*:*bad*, *intelligent*:*unintelligent*, *clever*:*stupid* sind Wortpaare, bei denen ein Teil einen wahren Komparativ, der andere aber nur einen Pseudokomparativ bilden kann. Daher unterscheiden sich die jeweiligen komparativen Sätze in ihren Implikationen. *Tom's quality is better than Jack's* impliziert nicht, dass Tom hoch qualifiziert ist und Jacks Qualifikation schlecht ist. Dagegen impliziert die Aussage *Jack's quality is worse than Tom's*, dass Jacks Qualifikation einigermaßen schlecht ist. Dies aber hindert wiederum nicht die logische Umwandlung zu *Tom's quality is better than Jack's*. Es ist deshalb einleuchtend, warum Lehrer (2002:500) diese Art der Relation als asymmetrische Antonymie bezeichnet. Außerdem ist zu erkennen, dass eines der Wortpaare eine positive und das andere eine negative Einschätzung darstellt und dass nur das positive Wort im „Wie-Fragetest“ relativ neutral verwendet werden kann.

Eine weitere Art der Opposition, die weder zur Antonymie noch zur Komplementarität gehört, ist die Konversion wie bei *teacher*:*student*, *husband*:*wife* und *buy*:*sell* (vgl. Cruse 1986:231-240, Lyons 1977:290-291). Konversion besteht dann, wenn ein Wort eines Wortpaares als Prädikat einer Proposition durch das andere ersetzt wird und bei gleichzeitigem Austausch der Argumente die ursprüngliche und die abgeleitete Proposition logisch äquivalent sind, wie in *A is a/the wife of B* und *B is a/the husband of A*. Die Sachlage kann komplizierter werden, wenn Konversion zwischen Verben besteht, die mehr als zwei Argumente benötigen:

> „Miriam gave a snuff-box to Arthur.
> Arthur received a snuff-box from Miriam.
> Harry sold the sarcophagus to the Emir.
> The Emir bought the sarcophagus from Harry.“ (Cruse 1986:233)

In den Beispielsätzen sind die obligatorischen und die fakultativen Argumente der dreistelligen Verben *give*:*receive* und *sell*:*buy* vertauscht. Cruse (1986: 234) nennt diese Art der Konversion „indirekte Konversion“.

<u>Meronymie</u>

Wenn Inklusion und Transitivität charakteristisch für Hierarchiebeziehungen sind (vgl. Murphy/Lasaline 1997:95), dann drückt neben der Hyponymie auch die Meronymie, die die semantische Teil-Ganzes-Relation darstellt, ein hierarchisches Begriffsverhältnis aus. Dies zeigt sich in der folgenden Definition:

> „X is a meronym of Y if and only if sentences of the form *A Y has Xs/an X* and *An X is a part of a Y* are normal when the noun phrases *an X*, *a Y* are interpreted generically.“(Cruse 1986:160)

Meronymie kann mit possessiven Konstruktionen ausgedrückt werden: *the wings of that bird* oder *the bird has wings*. Wenn ein Wort w_1 zu einem anderen Wort w_2 Meronym ist, dann ist w_2 Holonym zu w_1. Etwas in seine Bestandteile einzuteilen ist die grundlegende Analysemethode, die bei einem einfachen Objekt genauso wie bei komplexen Prozessen anwendbar ist (vgl. Pribbenow 2002:33-39). Die Schwierigkeit bei der Bestimmung der Meronymie ergibt sich aus dem Charakter des denotierten Sachverhalts und aus der daraus entstehenden Vielfältigkeit der Art und Weise, wie sich die Bestandteile eines Ganzen zusammensetzen können. Deswegen ist die Reichweite der Meronymie so groß, dass ihre jeweilige Art schwierig zu bestimmen ist. Winston/Chaffin/Herrmann (1987:417-444) unterscheiden sechs verschiedene Variationen der Meronymie: Komponente-Objekt (*branch*:*tree*), Mitglied-Gruppe *(tree*:*forest*), Portion-Menge (*slice*:*cake*), Material-Objekt (*aluminum*:*airplane*), Eigenschaft-Aktivität (*paying*:*shopping*) und Platz-Gebiet (*Princeton*:*New Jersey*).

Die Forschungen zur Meronymie beziehen sich zum großen Teil auf Nomen (vgl. Cruse 1986:157-178, Fellbaum 1998b:77). Dies liegt überwiegend daran, dass konkrete Gegenstände, auf die typischerweise von Nomen referiert wird, für eine Analyse besser geeignet sind als Handlungen oder Prozesse, auf die typischerweise von Verben referiert wird. Da Nomen prototypisch abgeschlossene physikalische Objekte denotieren, bei denen einzelne Teile relativ einfach vom zugehörigen Ganzen abgrenzbar sind, ist die Eigenschaft der Zerlegbarkeit auf Komponenten, Mitglieder und Materialien einzuschränken (vgl. Miller 1998:39). Eine Subhandlung bzw. ein Subprozess ist von anderen Subhandlungen und Prozessen nicht klar abtrennbar, und manchmal sind die Teile einer Handlung oder eines Prozesses nicht unbedingt selbst wieder Handlungen oder

Prozesse (vgl. Miller/Fellbaum 1992:218 f.). Obwohl Teilaktivität und zeitliche Inklusion – wie bei *buy*:*pay* und *sleep*:*snore* zu erkennen – normalerweise als gute Kriterien zur Feststellung von Meronymie dienen, weisen Miller/Fellbaum (1992:219) auf problematische Fälle wie *succeed*:*try* und *fatten*:*feed* hin, die anhand der Zeitinklusion schwer zu erklären sind. Eine mögliche Lösung wäre, eine Relation zwischen zwei Verben dann als Meronymie zu betrachten, wenn das eine das andere Verb lexikalisch impliziert, zeitlich umfasst und zwischen den beiden keine Troponymie besteht. Eine solche Relation ist in WordNet-2.0 als „Implikation" dargestellt. Wie oben in Abbildung 2-1 zu erkennen ist, ist der Begriff „Implikation" in der Implementierung von WordNet-2.0 enger definiert ist als bei Miller/Fellbaum (1992:222).

Transitivität ist, anders als bei der Hyperonymie, ein kritischer Punkt bei der Meronymie, worauf von vielen Linguisten hingewiesen wurde (vgl. Cruse 1986:165-168, Lyons 1977:322 f.). *Cuff* kann als Meronym sowohl von *sleeve* als auch von *jacket* betrachtet werden, wobei *sleeve* wiederum Meronym von *jacket* ist. Der Fall von *handle*:*house* ist hingegen nicht eindeutig, auch wenn Meronymie zwischen *handle*:*door* und *door*:*house* besteht. Cruse (1986:165 f.) erklärt diese Erscheinung mit dem Begriff „funktionale Domäne", der die funktionale Reichweite von Meronymen zu ihren Holonymen bezeichnet. Zum Beispiel hat ein Griff zwar die Funktion des Öffnens und Schließens, jedoch nur bei einer Tür, nicht aber beim ganzen Haus, während eine Stulpe die Funktion der Dekoration bei einem Ärmel, aber auch bei der ganzen Jacke haben kann. Hieran lässt sich zeigen, dass bei Wörtern, die der gleichen funktionalen Domäne angehören, Transitivität relativ problemlos möglich ist (vgl. Cruse 1986:165 f.). Ein weiterer interessanter Aspekt von Transitivität bei Meronymen ist ihre Fähigkeit, als Merkmale von Hyperonymen an deren Hyponyme vererbt werden zu können (vgl. Abschnitt 2.2.3.1). Das heißt, die Meronyme von w_1 können auch die Meronyme der Hyponyme von w_1 sein. Beispielsweise sind *beak* und *wing* Meronyme von *bird*, und *canary* ist Hyponym von *bird*. *Beak* und *wing* sind aber nicht nur die Meronyme von *bird*, sondern auch von *canary*. Miller (1998:38) plädiert dafür, Meronyme in der Hierarchie nicht zu hoch einzusetzen. Wenn man *wheel* als Meronym von *vehicle* betrachten würde, dann müsste *wheel* auch das von *sledge* sein, was aber nicht der Fall ist.

Das Problem resultiert daraus, dass das Wort *vehicle* als Hyperonym von *sledge* zu allgemein ist. Die Schwierigkeit besteht darin, dass es keine Wörter gibt, die Hyponyme von *vehicle* sind und sich nur darin voneinander unterscheiden, ob sie ein Rad besitzen oder nicht.

Kausale Relationen

Kausale Relationen können nicht nur zwischen Textteilen oder Sätzen, sondern auch zwischen einzelnen Wörtern bestehen (vgl. Nussbaumer 1991:188 ff.). Problematisch bei der Bestimmung der Kausalbeziehung zwischen Wörtern ist, dass die Berücksichtigung von kontextuellen Faktoren in manchen Fällen nicht vermeidbar ist, zum Beispiel stellt sich die Frage, ob die Relationen zwischen *bring* und *come* oder *have* und *give* kontextunabhängig als Ursache-Wirkung betrachtet werden können (vgl. Miller/Johnson-Laird 1976:468 f.) Die Bestimmung der Kausalbeziehung zwischen Wörtern ist davon abhängig, inwieweit kontextuelle Faktoren dabei berücksichtigt werden sollen. Es gibt jedoch Wortpaare, die relativ kontextunabhängig kausale Relation darstellen, wie *kill*:*die*, oder *persuade*:*believe*. Da die Ursache-Wirkung-Relation sich typischerweise auf die Veränderung von Prozessen, Handlungen oder Zuständen bezieht, ist es nicht besonders verwunderlich, dass sie normalerweise zwischen Verben zu finden ist. Im Englischen wird die Ursache oft durch Derivation ausgedrückt. Beispiele sind von Adjektiven abgeleitete Verben wie *modernize* oder *enrich* und von Nomen abgeleite wie *capitalize* und *encourage*.

Derivation

Während eine Komposition mindestens ein zusätzliches freies Morphem benötigt, geschieht die Derivation durch Affixe, die keine selbstständige lexikalische Bedeutung besitzen, zum Beispiel *arrive*:*arrival*, *pay*:*payment*, *grade*:*degrade* (vgl. Lipka 2002:100 ff., Schippan 2002:43 f., 114-117). Durch die Derivation kann sich nicht nur die Wortklasse verändern, wie im Fall von *pay*:*payment*, sondern es kann auch eine semantische Veränderung erfolgen wie bei *grade*:*degrade*, wo der Zusammenhang zwischen semantischen und morphologischen Aspekten deutlich zu erkennen ist. Im praktischen Teil dieser Arbeit wird v. a. die Nominalisierung im Mittelpunkt stehen, also diejenige Form der Derivation, die ohne wesentliche semantische Veränderung Verben

zu Nomen macht. Im Englischen erfolgt dies hauptsächlich durch Suffigierung wie bei *activate*:*activation*, *employ*:*employment*, *close*:*closure*, *dispense*: *dispensary*, *kill*:*killer* etc. (vgl. Bauer 1983:221 f.). Als Sonderfall gilt die sog. „Null-Derivation“, die oft als „Konversion“ bezeichnet wird und bei der ein Wortartwechsel ohne morphologische Veränderung erfolgt, wie bei *cook*(n):*cook*(v) und *fight*(n):*fight*(v) (vgl. Cruse 2000:149 f., Lipka 2002:101).

2.2 Lexikalische Relationen und mentales Lexikon

2.2.1 Lexikalische Relationen und die Struktur des mentalen Lexikons

Da Konzepte als Entitäten betrachtet werden, die zur mentalen Ebene gehören, behauptet die Annahme, dass Konzepte auf der mentalen Ebene netzwerkartig strukturiert sind, ihre Gültigkeit unabhängig von der sprachlichen Realisierung der Konzepte. Wenn man die Repräsentation der Wortkonzepte betrachtet, so ist oft vom „mentalen Lexikon" die Rede, womit die Organisation der den Wortformen zugewiesenen Konzepte auf der mentalen Ebene bezeichnet wird (vgl. Lipka 2002:197 ff., Aitchison 1994:105-125). Viele Arbeiten zum mentalen Lexikon gehen von der Annahme aus, dass Konzepte netzwerkartig strukturiert sind: Zum Beispiel fasst Fodor (1983:80) den Begriff „mentales Lexikon" als „a sort of connected graph, with items at the nodes and with paths from each item to several others" auf. Ein grundlegendes Problem ist die Frage, welche und wie viele Informationen im mentalen Lexikon repräsentiert sein sollen:

> „The question of how ambitious a lexicon should be, that is, in how much detail the meaning of word concept should be represented in the lexicon, can not be answered simply." (Kintsch 1974:19)

Im Prinzip kann ein Wort mit seinen Verbindungen zu allen anderen Wörtern in einer Netzwerkstruktur gespeichert werden. In diesem Zusammenhang kann sogar das sog. propositionale Netzwerk, das üblicherweise episodisches Wissen darstellt (vgl. Anderson 1995a:144-147, Stillings et al. 1987:145-150), als ein Modell für das mentale Lexikon betrachtet werden (vgl. Abbildung 2-2).[22] Setzt man für das mentale Lexikon ein Netzwerkmodell an, dann ist die entscheidende Frage, welche Relationen zwischen den Konzepten bestehen, die das mentale Lexikon konstituieren.

[22] Unter mathematischer Hinsicht kann man bei der Betrachtung der Netzwerkmodelle die Graphentheorie heranziehen (vgl. Dietze 1994:55 f., Klix 1988, Sowa 1984). Wie die Abbildung 2-2 zeigt, in der die durch Pfeile dargestellten Kanten die Relationen zwischen den Knoten bezeichnen, können die Netzwerkmodelle als ein gerichteter Graph erfasst werden. Obwohl die Pfeile bei manchen Netzwerkmodellen nicht explizit dargestellt werden, kann man sie als gerichtet betrachten, wenn die Informationen über die Relationen explizit angegeben sind. Der gerichtete Graph ist eine Voraussetzung für die logische Analyse der Kanteneigenschaften, zu denen Transitivität, Reflexivität und Symmetrie gehören. Diese logische Charakteristik ist nicht nur für die Untersuchung der semantischen Eigenschaften der lexikalischen Relationen relevant, sondern auch für den Entwurf des Algorithmus, durch den die lexikalischen Relationen in Texten erkannt werden sollen. Darauf wird in Abschnitt 3.3.3 genauer eingegangen.

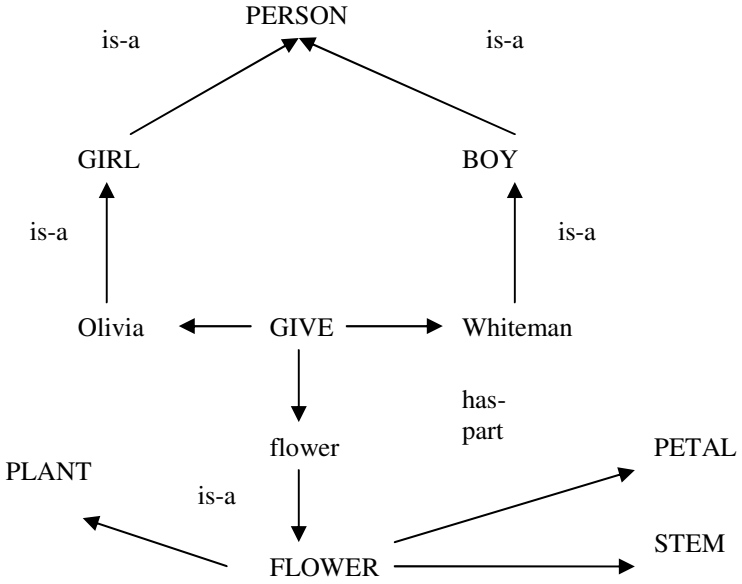


Abbildung 2-2: Netzwerkmodell von Stillings et al. (1987:147)

Das Bestimmungsproblem ergibt sich auch beim Spreading-Activity-Modell von Collins/Loftus (1975), deren Interesse der Aktivierungsausbreitung von Konzepten im neuronalen Netzwerk gilt. Die gelernten Konzepte haben je nach dem Verlauf des Lernprozesses unter sich eine verschiedene Verbindungsstärke, die durch die unterschiedliche Länge der Kanten in Abbildung 2-3 gekennzeichnet ist. Die Verbindungsstärke kann auch als der Grad der Aktivierung betrachtet wird, die durch die inneren und äußeren Stimuli ausgelöst wird.[23]

[23] Dieses Modell der Konzeptaktivierung, das hinsichtlich der angenommenen Netzwerkstruktur mit der Schematheorie vereinbar ist, kann auch auf Textproduktion und Textverstehen übertragen werden. Obwohl sich die Aufmerksamkeit des Textproduzenten besonders stark auf den Teil der konzeptuellen Struktur richtet, der für die Versprachlichung zuständig ist, werden auch Konzepte in geringerer Stärke mitaktiviert, die mit den Wortkonzepten assoziiert sind. Auf der Seite des Textrezipienten werden die mit den sprachlichen Ausdrücken direkt verbundenen Konzepte aktiviert, aber auch mit den Konzepten assoziierte Konzepte werden aktiviert, sodass das Kontextwissen aktiviert wird und damit der Verstehensvorgang vollzogen werden kann (vgl. Abschnitte 2.3.3 und 2.3.4).

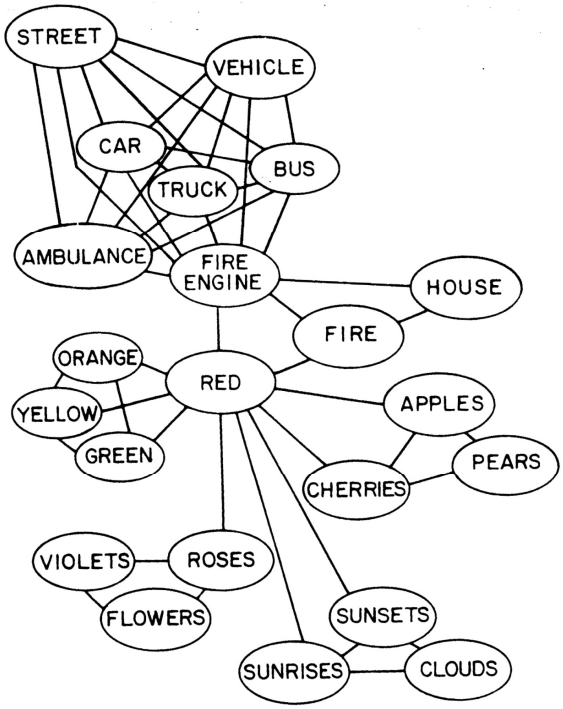


Abbildung 2-3: Das Speading-Activity-Modell von Collins und Loftus (1975:412)

Die grundsätzliche Frage bei der Bestimmung des mentalen Lexikons kann hinsichtlich der Aktivierungsausbreitung so formuliert werden: Welche Art von Relationen sind verbindungsstark und welche sind verbindungsschwach? Aitchison (1994:105-125) nennt in diesem Zusammenhang drei Relationsarten, die für die Struktur des mentalen Lexikon relevant sind: Erstens geht es um Wörter, die zu demselben thematischen Feld gehören. Zum Beispiel werden die Wörter *thread*, *pins* und *eyes* häufig als mit *needle* assoziierte Wörter genannt.[24] Zweitens geht es um Wörter, die durch lexikalische Relationen miteinander verbunden sind, wobei die Verbindung bei der Antonymie besonders stark ist. Dass thematische und semantische Bereiche hierbei nicht sauber zu trennen sind, ist auch dadurch erkennbar, dass die Wörter, die zu demselben thematischen Feld gehören, oft durch eine lexikalische Relation wie Meronymie oder durch die Kombination mehrerer lexikalischer Relationen dargestellt werden können. Was Aitchison zusätzlich betont, ist die Verbindung der Wör-

[24] Warum Wörter eines thematischen Feldes stark miteinander verknüpft sind, kann am besten mit der Schematheorie erklärt werden, die das Erfahrungslernen auf den kulturellen und sozialen Kontext zurückführt.

ter, die zu derselben Wortklasse gehören.[25] Da bei lexikalischen Relationen vorausgesetzt ist, dass die Wörter zu derselben Wortklasse gehören, lässt sich sagen, dass die Verbindungsstärke zwischen diesen Wörtern höher ist als zwischen anderen Wörtern derselben Wortklasse.

Das Modell des semantischen Gedächtnisses, in dem den lexikalischen Relationen eine besondere Bedeutung zukommt, ist das von Collins und Quillian (1969):

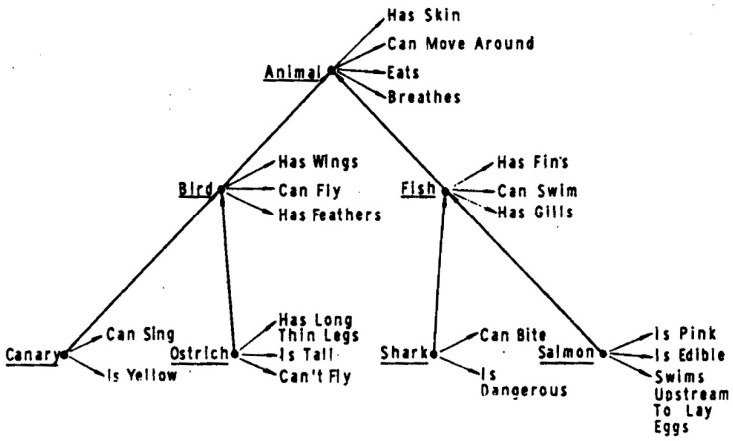


Abbildung 2-4: Hierarchische Struktur des semantischen Gedächtnisses (Collins/ Quillian 1969:241)

Abbildung 2-4 zeigt, dass die semantischen Informationen vor allem durch die Klasseninklusion und die Angabe der Eigenschaften von Konzepten repräsentiert sind. Außerdem ist zu sehen, dass die Beziehung zwischen Konzepten und einigen ihrer Eigenschaften wie HAS-SKIN, HAS-WINGS etc. meronymen Charakter zeigt. Auch das Modell von Collins und Quillan zeigt, dass Hyponymie und Meronymie als fundamentale Relationen für die Struktur des mentalen Lexikons gelten, obwohl diese Annahme nicht völlig der psychologischen Realität entspricht (vgl. Abschnitt 2.2.3.1). Die Vernetzung von Konzepten durch Hyponymie und Meronymie ist allerdings viel komplizierter strukturiert, als in Abbildung 2-4 dargestellt ist. Zum Beispiel können Wörter sowohl Meronyme als auch Hyponyme sein. *Beak*, *bill* und *neb* sind etwa Meronyme von

[25] Dies kann sowohl auf syntaktische als auch auf semantische Gründe zurückgeführt werden. Typischerweise bezeichnen Nomen Objekte und Verben Handlungen, Prozesse und Zustände. Deshalb haben die Wörter, die zu einer Wortklasse gehören, oft einen gemeinsamen semantischen Charakter, der weiter bei der syntaktischen Selektion eine Rolle spielt.

bird und gleichzeitig Hyponyme von *jaw*, das wiederum ein Meronym von *skull* und ein Hyponym von *skletal_structures* ist (vgl. Miller 1998:38). Noch komplizierter wird es, worauf Abbildung 2-5 hindeutet, wenn man berücksichtigt, dass je nach den Eigenschaften der Konzepte unterschiedliche Konfigurationen durch Hyponymie und Meronymie entstehen (vgl. Rips/Estin 1998):

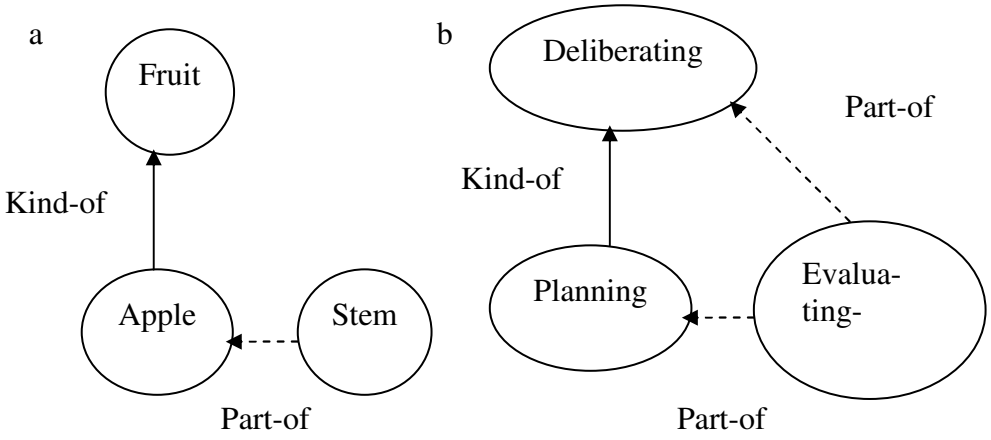


Abbildung 2-5: Hyponymie und Meronymie zwischen Objekt- und Aktivitätskonzepten (Rips/Estin 1998:312).

2.2.2 Lexikalische Relationen als Resultat von Informationsreduktion in der konzeptuellen Struktur

Die Tatsache, dass lexikalische Relationen die Struktur des mentalen Lexikons entscheidend bestimmen, ist ein wichtiges Argument für die Implementierung von WordNet, bei dem das semantische Gedächtnis vor allem durch lexikalische Relationen repräsentiert wird. Eine mögliche Antwort auf die Frage, warum lexikalische Relationen einen so wichtigen Teil des mentalen Lexikon bilden, kann man unter Verweis auf die zwei grundlegenden kognitiven Fähigkeiten der Kategorisierung bzw. Klassifikation und der Dekomposition in Subteile beantworten (vgl. Pribbenow 2002:36). Aus zwei Gründen kann man Meronymie als die Relation betrachten, die auf noch grundlegenderen kognitiven Prozessen als Klassifikation bzw. Kategorisierung beruht: Erstens zeigen Forschungen, dass Meronymie früher als Hyponymie erworben wird und sogar dreijährige Kinder in der Lage sind, Meronymie zu erkennen (vgl. Pribbenow

2002:37). Zweitens ist sie nicht nur Voraussetzung für die Analyse von relativ komplexen Sachverhalten, sondern auch für die Erkennung von sehr primitiven Objekten, deren Konzepte noch nicht in der taxonomischen Struktur etabliert sind. Weiter sind Menschen dazu fähig, nach den dominanten gemeinsamen Informationen und individuellen Konzepten eine neue Klasse oder Kategorie zu bilden (vgl. Anderson 1995a:352-358). Marvis/Pani (1980:496-522) ist es sogar gelungen, fünfjährigen Kindern die Klassifikation künstlicher Objekte beizubringen. Diese Fähigkeit kommt besonders beim Erwerb neuer Wortkonzepte zum Tragen, wobei hier Taxonomie eine wichtige Rolle spielt,[26] die als eine besondere Art der Hyponymie gelten kann (vgl. Abschnitt 2.1). Es ist auch eine gängige Annahme in der kognitiven Wissenschaft, dass Taxonomie eine sehr informationsreduzierte Wissensstruktur darstellt (vgl. Jackendoff 1983:143).

Wissenschaftler haben darauf hingewiesen, dass lexikalische Relationen auf einer noch weitergehenden Informationsreduktion auf der konzeptuellen Ebene[27] basieren: Nach Minsky (1981:102) sind Menschen in der Lage, aufgenommene

[26] Rothweiler/Meibauer (1999: 21) listen folgende Annahmen für den Konzepterwerb auf, zu denen auch die Taxonomie-Annahme gehört:
„– die Annahme, dass sich Wörter auf Klassen und nicht auf Individuen beziehen (*type assumption*, Clark 1993);
– die Annahme, dass alle Wörter im Lexikon einer einzigen Hierarchieebene angehören (*single level assumption*) (Clark 1993);
– die Annahme, dass sich Wörter auf ganze Objekte beziehen und nicht nur auf einen Teil oder eine Eigenschaft (*whole object assumption*, Markman 1989; Mervis 1987; *object scope principle*, Golinkoff et al. 1994);
– die Taxonomie-Annahme, der zufolge sich Wörter auf taxonomisch organisierte Kategorien beziehen (*taxonomic assumption*, Marman & Hutchison 1984; *categorial scope principle*, Golinkoff et al. 1994);
– die Annahme, dass sich ein neues Wort auf eine bisher unbekannte Kategorie bezieht (*novel name – nameless category principle*, Golinkoff et al. 1994)“

[27] Obwohl kognitive Prozesse kontinuierlichen Charakter besitzen und daher nur schwer klare Prozessstufen unterschieden werden können, wird oft der Versuch unternommen, die Repräsentationsebene zu differenzieren. Es ist in der kognitiven Wissenschaft üblich, die mentale Ebene in die konzeptuelle und die sog. referenzielle Ebene einzuteilen (vgl. z. B. Jackendoff 1983:16 ff., 35 ff., Schwarz 1992:90-97, Sucharowsky 1996:163-166). Die referenzielle Ebene, die in der traditionellen Linguistik mit der realen Welt identifiziert wurde und deswegen unmittelbar mit perzeptiven Informationen zusammenhängt, verlagert sich bei Jackendoff auf die mentale Ebene, auf welche die reale Außenwelt projiziert wird (vgl. Jackendoff 1983: 23-29). Eine ähnliche Ansicht vertreten auch Schwarz und Lakoff, die jeweils propositionale und analoge Repräsentationen bzw. ein propositionales und ein bildschematisches Modell unterscheiden (vgl. Lakoff 1987:113 f., Schwarz 1992:19). Wenn man davon ausgeht, dass Konzepte die projizierten Sachverhalte der Außenwelt auf der konzeptuellen Ebene darstellen (vgl. Abschnitt 1.3.1), dann impliziert Konzeptbildung auch Informationsreduktion. Der Zusammenhang zwischen Repräsentation und Informationsreduktion wird auch durch das folgende Zitat von Palmer (1978:262) deutlich: „Representation is […] something that stands for something else. […] This description implies the existence of two related but functionally separate worlds: the represented world and the representing world. The job of the representing world is to reflect some aspects of the represented world in some fashion.“

Informationen zu kondensieren und zu konventionalisieren und diese Informationen Wörtern zuzuweisen. Wierzbicka (1996) betrachtet die Prädikate PART (OF), KIND (OF), BECAUSE, THE SAME, OTHER und NOT, die den lexikalischen Relationen entsprechen, sogar als semantische Primitive. Die Auffassung, dass lexikalische Relationen einen hoch abstrakten Teil der konzeptuellen Struktur darstellen, wird vor allem von Anhängern der Modularitätsannahme vertreten, die eine konzeptuelle und sprachliche Ebene voneinander unterscheiden (vgl. Abschnitt 1.3.3). Zum Beispiel meint Blank (2001:131-135), dass „die sememische Bedeutung", die für das einzelsprachliche System relevant ist, aus dem enzyklopädischen Wissen abstrahiert ist. Nach Dietze (1994:24-44) beruhen die lexikalischen Relationen wie Hyperonymie, Synonymie und Antonymie auf dem sog. „Sem" (semantisches Merkmal), das als „abstrahiertes Wissenselement" gilt und dessen typische Eigenschaft die Invarianz ist.[28]

Die beiden Annahmen, dass lexikalische Relationen den sehr kompakten und hoch abstrahierten Teil einer Wissensstruktur repräsentieren und dass das mentale Lexikon so wie andere konzeptuelle Strukturen netzwerkartig strukturiert ist, dienen als wichtige Begründung für den Einsatz von WordNet für die automatische Indexierung. Das Wissen, auf dem lexikalische Relationen basieren und das deswegen als sprachliches Wissen gelten kann, unterscheidet sich nur wenig zwischen Individuen, die dieselbe Sprache sprechen, und es bleibt relativ konstant. Da die lexikalischen Relationen einen relativ konstanten und kontextunabhängigen Wissensbereich repräsentieren, sind sie eine geeignete Wissensbasis für die Textanalyse.

[28] Wenn man Wortkonzepte wie andere Konzepte als netzwerkartig strukturiert auffasst, dann lassen sich die durch die lexikalischen Relationen verknüpften Teile der konzeptuellen Struktur als Subgraph des Graphs erfassen, der die ganze Wissensstruktur darstellt, die Menschen zur Verfügung haben. In der Sprache der Graphentheorie kann Reduktion als Reduzierung der Kanten erfasst werden (vgl. Klix 1988:24). Je höher die Anzahl der Knoten ist, desto höher ist der Reduktionsgrad. Weiter kann die Granularität der Repräsentation, die als ein wichtiges Kriterium für die Unterscheidung von Netzwerkmodellen dient, als die Anzahl der Kanten aufgefasst werden, die das ganze Netzwerk bilden.

2.2.3 WordNet

2.2.3.1 Das Problem der hierarchischen Organisation des semantischen Gedächtnisses

Dass Hyponymie hierarchischen Charakter besitzt und ihre Eigenschaften teilweise durch Meronyme dargestellt werden können, führt zur Frage nach der Struktur des mentalen Lexikons. Wie das Modell von Collins und Quillian zeigt (vgl. Abschnitt 2.2.1), kann eine Eigenschaft wie HAT-FLÜGEL von VOGEL durch die Meronymie *wing*:*bird* dargestellt werden. Es zeigt sich, dass die Meronymie sich auch auf Hyponyme wie *canary* und *ostrich* vererbt. Collins/Quillian (1969:242-246) haben experimentell untersucht, ob die hierarchische Relation im menschlichen Gedächtnis die dominante Relation ist. Versuchspersonen wurden dazu angehalten, so schnell wie möglich die Sätze *a canary can sing*, *a canary can fly* und *a canary has skin* als richtig oder falsch zu beurteilen, wobei die für jedes Urteil benötigte Zeit gemessen wurde. Ihre Erwartung, dass die Sätze *a canary can fly* und *a canary has skin* aufgrund der Eigenschaften FLIEGEN-KÖNNEN und HAT-HAUT, die den höher angesiedelten Konzepten VOGEL und TIER zugeordnet sind, bei der Entscheidung mehr Zeit als der Satz *a canary can sing* benötigen, bei dem die Eigenschaft SING direkt dem Konzept CANARY zugeordnet ist, kann durch ihr Experiment bestätigt werden. Sogar der Satz *a canary is an animal* kostet mehr Zeit als *a canary is a bird*, der wiederum mehr Zeit als der Satz *a canary is a canary* kostet. Aber diese Beobachtung kann nicht direkt zu der Schlussfolgerung führen, dass Wortkonzepte strikt hierarchisch organisiert sind, da das Ergebnis auch aus anderen Gründen wie etwa der Häufigkeit des gemeinsamen Vorkommens im Alltag verursacht sein kann. Weitere Untersuchungen ergaben, dass die Verbindung von Konzepten wie etwa NAIL und HAMMER wegen ihrer gemeinsamen Vorkommenshäufigkeit genau so stark sein kann wie bei den Konzepten APFEL und OBST, die einer unmittelbaren konzeptuellen Hierarchie unterliegen (vgl. Aitchison 1994:l20 f.). Es konnte sogar gezeigt werden, dass Merkmale nicht unbedingt nur den Konzepten der höchsten Ebene zugeordnet sind, sondern auf verschiedenen hierarchischen Ebenen mehrfach abgespeichert sein können (vgl. Collins/Quillian 1969:242). Die Existenz von Re-

dundanzen im Netzwerk zeigt also, dass hierarchische Relationen nicht unbedingt die dominanten Relationen im Gedächtnis sind und dass das ökonomische Prinzip nicht unbedingt das oberste Prinzip für die Strukturierung des Gedächtnisses ist. Dennoch werden hierarchische Relationen nicht nur von Informatikern, sondern auch von Lexikographen oft beim Entwurf und der Implementierung von Wissensrepräsentation ins Zentrum gestellt, wie dies auch bei WordNet der Fall ist (vgl. Miller 1998:31 ff.). Dies liegt hauptsächlich daran, dass dadurch die Durchführung der Inferenz für Merkmale, die der höheren Ebene zugeordnet sind, und die Berücksichtigung der Speicherökonomie erleichtert werden (vgl. Habel 1985:453). Außerdem verweist Miller (1998:31 ff.) darauf, dass Hierarchie immerhin ein dominantes Prinzip für die Organisation der Nomen auf der mentalen Ebene ist.

Wissensrepräsentation ist zweckgebunden. Das heißt, man kann nicht alle Aspekte des Wissens auf einmal darstellen, sondern nur intendierte. Dies betrifft auch WordNet, das zum Beispiel die thematischen Relationen, welche Konzepte auf der mentalen Ebene so eng zusammenbinden wie die lexikalischen Relationen, nicht vollständig in Betracht zieht. Problematisch wäre diese Art der Relationen für den Zweck dieser Arbeit, da diese manchmal keinen expliziten Relationscharakter haben (vgl. Abschnitt 2.3.2). Im Gegensatz dazu haben lexikalische Relationen relativ konstante semantische Eigenschaften, und dadurch können sie, wenn die Relationen in der Wissensbasis explizit bezeichnet sind, für weitere informationsverarbeitende Prozesse genutzt werden. Das eigentlich ist der grundlegende Nutzen bei der Verwendung von WordNet.

2.2.3.2 Dateistruktur

Im Rahmen von WordNet sind ungefähr 150.000 Wörter – Nomen, Verben, Adjektive und Adverbien – je nach Wortklasse in getrennten Dateien eingetragen, wobei, anders als bei konventionellen Wörterbüchern, keine Informationen über Aussprache und Flexion der Wörter abgespeichert sind. Die getrennte Speicherung nach Wortklasse kann durch die in Abschnitt 2.2.1 beschriebene psycholinguistische Begründung legitimiert werden, dass Wörter derselben Wortklasse jeweils enger als mit Wörtern anderer Wortklassen verbunden sind.

Für jede Wortklasse existiert eine Indexdatei und eine Datendatei, deren Format folgendermaßen aussieht:

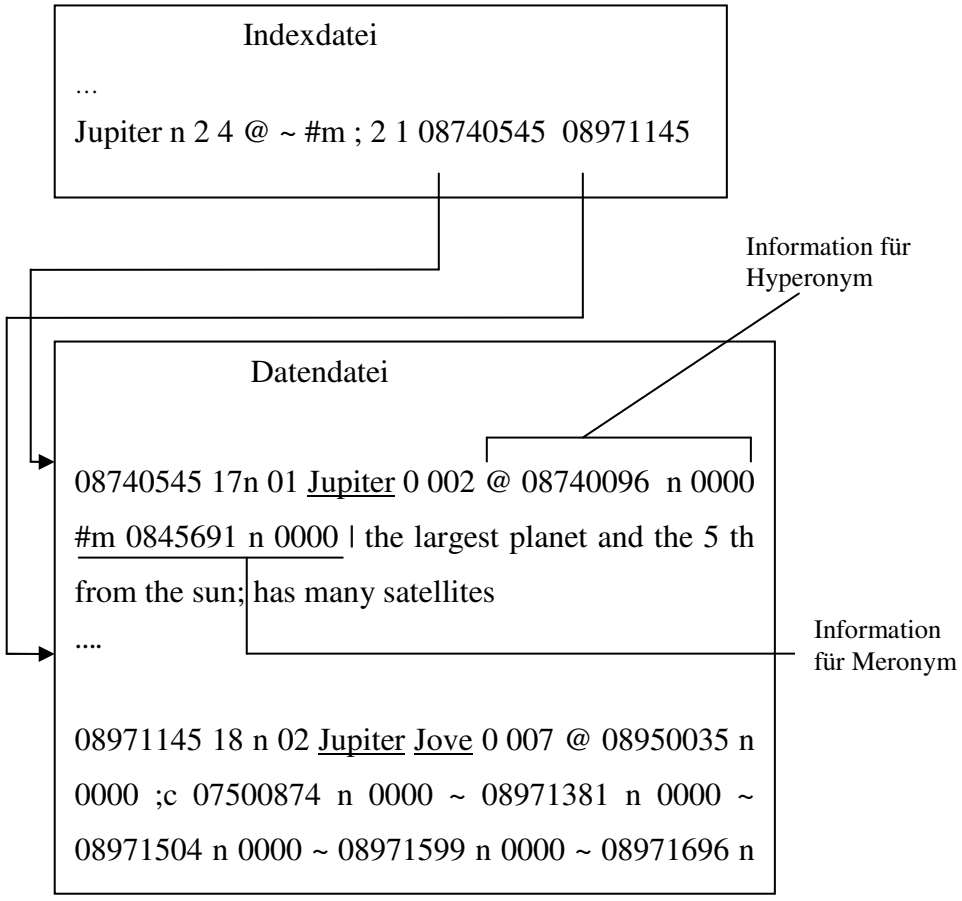


Abbildung 2-6: Dateistruktur von WordNet

Ein Eintrag der Index-Datei entspricht einem Lemma, also einer Wortform mit möglichen Bedeutungen bzw. Konzepten, deren Informationen in der Daten-Datei repräsentiert sind. Dies wird dadurch ermöglicht, dass die Dateiadressen, die die Anfangsstellen jedes Eintrags in der Daten-Datei bezeichnen, in der Indexdatei als Referenz auf die möglichen Belegformen angegeben sind. Im obigen Fall hat das Lemma <Jupiter> zwei mögliche Belegformen, auf die jeweils die Dateiadressen 08740545 bzw. 08971145 zeigen. Die Repräsentation eines Konzepts ist hauptsächlich durch das Synset realisiert, in dem die Gruppe der Synonyme enthalten ist, wobei die Mitglieder eines Synsets dasselbe Konzept darstellen. Ein Konzept wird also durch die Menge der Wörter repräsentiert,

die unter Synonymie stehen.[29] Ein Synset ist in der obigen Abbildung durch die unterstrichenen Teile in der Daten-Datei dargestellt. Die erste Belegform von „Jupiter“ hat nur ein Mitglied des Synsets {*Jupiter*}, das den „Jupiter“ heißenden Planet darstellt, für die zweite ist aber durch *jove* klarer, dass das Synset {*Jupiter*, *Jove*} das Konzept darstellt, bei dem es sich um einen römischen Gott handelt. Die weiteren lexikalischen Relationen unterliegen in der Daten-Datei dem folgenden Format, wobei „Source-Synset“ das aktuelle Synset bezeichnet und „Ziel-Synset“ das Synset, das mit dem aktuellen Synset in einer lexikalischen Relation steht:

> *Symbole für lexikalische Relationen*[30] *Die Dateiadresse des Ziel-Synsets Wortklasse des Ziel-Synsets Angabe der Datei, in der das Target-Synset gespeichert ist*

Da ein Synset mehr als eine lexikalische Relation besitzen kann, kann dieses Format bei der Darstellung jeder Relation zum Einsatz kommen. Zum Beispiel hat *Jupiter* im Sinne von „Planet“ ein Hyperonym und ein Holonym.

[29] Dies wird von Miller und Fellbaum (1992:200 f.) wie folgt begründet: „If the person who reads the definition needs merely to identify a familiar concept, then a synonym (or near synonym) is often sufficient. For example, someone who knows that *board* can mean either a piece of lumber or a group of people assembled for some purpose will be able to pick out the intended meaning with no more than *plank* or *committee*.“

[30] Zum Beispiel steht @ für Hyperonym und #m für die Mitglied-Gruppe-Relation. Zu den weiteren Zeichen für die Relationen s. WordNet 3.0 Reference Manual. In: http://wordnet.princeton.edu/man2.0/2.0/wninput.5WN.html (09. 08. 2006).

2.3 Lexikalische Relationen und Textverarbeitung

2.3.1 Kohäsion und Kohärenz

Das Wort „Text“ kommt ursprünglich von dem lateinischen Wort „textus“, das eigentlich „Gewebe“ bedeutet, aber metaphorisch auf die „Verwebung“ oder „Verkettung“ eines Werkes hinweist (vgl. Ehlich 1984:10). Die „Verkettung“ in Texten – im Sinne von sprachlichen Kommunikationseinheiten – bezeichnet man in der Textlinguistik meist mit den Begriffen „Kohäsion“ und „Kohärenz“, die von den Wissenschaftlern allerdings uneinheitlich verwendet werden. Dies liegt vor allem an den unterschiedlichen Sprachmodellen, mit denen Textforschung betrieben wird: Unter dem Einfluss der strukturellen Linguistik und der generativen Transformationsgrammatik resultiert die Textbetrachtung erst aus dem Versuch der Überwindung der Probleme, die aus der am einzelnen Satz orientierten Forschung entstehen. Deswegen standen zuerst die satzübergreifenden, textkonstituierenden linguistischen Elemente im Mittelpunkt, etwa die Satzverknüpfung durch Pronomina (pronominale Verkettung), andere Proformen und Konjunktionen (vgl. z. B. Harweg 1968, Isenberg 1970, Steinitz 1968). Da in dieser Forschungsrichtung davon ausgegangen wird, dass das Sprachsystem seine eigene Autonomie besitzt und sich Textualität hauptsächlich auf der grammatischen Ebene äußert, ist eine klare Unterscheidung zwischen Kohärenz und Kohäsion möglich. Kohäsion bezeichnet hiernach die hauptsächlich grammatischen Relationen, Kohärenz bezieht sich auf den Sinnzusammenhang, der außerhalb des Sprachsystems liegt (vgl. Rickheit/Schade 2000:276-279). Die klare Unterscheidung zwischen Kohäsion und Kohärenz wird erst problematisch, wenn die explizite oder implizite Trennung von sprachlichem Wissen und Weltwissen bestritten und deshalb im Sprachmodell keine eigene sprachspezifische Domäne vorgesehen wird. Halliday/Hasan (1976) unterscheiden beispielsweise in ihrer grundlegenden Arbeit nicht zwischen Kohäsion und Kohärenz. „Kohäsion“ – die für sie auf dem „context of situation“[31] beruht – bezieht sich bei ihnen in einem sehr allgemeinen Sinne auf

[31] Halliday/Hasan (1976:22) beschreiben „context of situation“ zusammenfassend wie folgt: „The FIELD is the total event, in which the text is functioning, together with the purpositive activity of the speaker – or writer; it thus includes the subject-matter as one element in it. The MODE is the function of the text in the event, including therefore both the channel taken by the language – spoken or written, ex tempore or prepared – and its genre, or rhetorical mode, as

die Bedeutungsbeziehungen innerhalb eines Textes, die diesen als Text definieren („relations of meaning that exist within the text, and that define it as text“) und die auch durch sprachliche Mittel ausgedrückt werden können (vgl. Halliday/Hasan 1976:4). Eine ähnliche Auffassung vertritt auch Brinker (2005:18), aber er verwendet anstelle von „Kohäsion“ den Terminus „Kohärenz“. Die Schwierigkeit der Differenzierung der beiden Begriffe betrifft besonders den kognitiven Ansatz in der semantischen Forschung, bei dem kein Unterschied zwischen sprachlichem und Weltwissen gemacht wird (vgl. Abschnitt 1.3.3). Dies führt dort zu der Ansicht, dass Kohärenz nicht mehr als inhärente Eigenschaft von Texten, sondern als eine Eigenschaft der menschlichen kognitiven Funktion aufzufassen ist (vgl. Gernsbach/Givón 1995a:XI). Weiter unterscheidet Givón (1995) Kohärenz auf der Textebene und auf der mentalen Ebene: Während sie auf der Textebene nur als „methodologically useful observable artifact“ zu betrachten ist, stellt sie sich auf der mentalen Ebene als der Verknüpfungsprozess von Konzept und Informationen dar (vgl. Givón 1995a:61-65). Dass Givón für die Bestimmung von Kohärenz so stark das Konstruktionsmoment betont, liegt vermutlich an seinem bevorzugten Forschungsgegenstand, der gesprochenen Sprache, die stark prozedurale Züge aufweist. Von Interesse für die automatische Textanalyse ist jedoch nur die Kohärenz auf der Textebene, da nur hier die sprachlichen Ausdrücke (Wortformen) zu finden sind und als physikalische Zugriffseinheiten für die maschinelle Verarbeitung dienen. Es ist erkennbar, dass die begriffliche Unterscheidung von Kohäsion und Kohärenz auch mit dem Problem der Schnittstelle zwischen semantischer und konzeptueller Ebene zusammenhängt (vgl. Abschnitt 1.3.3). In dieser Arbeit wird die lexikalische Kohäsion über die kontextunabhängig erkennbaren lexikalischen Relationen im Text erfasst, und Kohärenz wird als der Sinnzusammenhang verstanden, der durch die Konzepte, die oft nicht kontextunabhängig interpretierbar sind, hergestellt wird. Diese Annahme ist vereinbar mit Halliday/Hasan (1976:285), die die Realisierung der lexikalischen Kohäsion vor allem auf das lexikalische System zurückführen.

---

narrative, didactic, persuasive, ‘phatic communication’ and so on. The TENOR refers to the type of role interaction, the set of relevant social relations, permanent and temporary, among the participants involved. Field, mode and tenor collectively define the context of situation of a text […].The linguistic features which are typically associated with a configuration of contextual features – with particular values of the field, mode and tenor – constitute a register. “

2.3.2 Lexikalische Kohäsion

Welche Beziehungen zwischen Einheiten des Textes zur Kohäsion gerechnet werden sollen, ist unter Linguisten umstritten. Zum Beispiel unterscheidet Isenberg (1977, 1977a), der eine der umfangreichsten Klassifikationen der Kohäsionsbeziehungen aufgestellt hat, die Kohäsion, die einzelne Satzglieder anspricht, etwa durch Pronomina, Artikel oder Satzkonnektoren, von derjenigen, die sich durch Wortstellung oder Prosodie auf den gesamten Satz bezieht. Halliday/Hasan (1976) unterteilen die Kohäsion in grammatische und lexikalische Kohäsion, wobei Referenz, Substitution und Ellipse Ersterer zugeordnet werden und die durch Vollwörter realisierte Verbindung der Letzteren. Halliday/Hasan (1976:274-278) unterscheiden ebenfalls zwei Arten der lexikalischen Kohäsion, und zwar „Reiteration“ und „Kollokation“: Reiteration (lexikalische Wiederaufnahme) ist eine Koreferenzbeziehung zwischen zwei Wörtern, die Wiederholung, Synonymie und Hyperonymie umfasst. Kollokation hingegen umfasst Wortpaare, die gehäuft zusammen auftreten, aber nicht in einer Koreferenzbeziehung zueinander stehen. Was bei dieser Klassifikation zu Verwirrung führen kann, ist die von Halliday/Hasan vorgenommene Zuweisung von Meronymie und Antonymie zur Kollokation, da man unter dieser in der Linguistik normalerweise nur die syntagmatische Relation zwischen Wortpaaren wie *moon*:*space shuttle* und *ill*:*doctor* versteht, deren Auftretenshäufigkeit den Zufall übersteigt (vgl. Hoey 1991:7), die aber nicht durch die üblichen lexikalischen Relationen miteinander verbunden sind. Das gemeinsame Vorkommen beruht sicherlich auf semantischen Ähnlichkeiten, aber das lässt sich nicht ganz systematisch und einheitlich erklären (vgl. Lyons 1977:220-226). Man kann versuchen, Kollokationen durch statistische Verfahren[32] zu ermitteln. Problematisch wird dabei für diese Arbeit, dass Informationen über die Eigenschaft solcher Relationen auf diesem Weg nicht zu ermitteln sind. Ein anderes Problem ergibt sich bei der Abgrenzung der Reichweite von Kollokation, so argumentiert etwa Hasan (1984:195) gegen ihre Berücksichtigung bei der Betrachtung der lexikalischen Kohäsion:

[32] Zur statistisch basierten Ermittlung der Assoziationsstärke s. Manning/Schütze 2002:162-177.

„While I firmly believe that behind the notion of collocation is an intuitive reality, I have come to accept the fact that unless we can unpack the details of the relations involved in collocation in the Firthian sense, it is best to avoid the category in research."

Aus diesem Grund schlägt Hasan (1984:202) die folgenden Arten der lexikalischen Kohäsion vor:

„Categories of lexical cohesion
A: General
i. repetition *leave, leaving, left*
ii. synonymy *leave, depart*
iii. antonymy *leave, arrive*
iv. hyponymy *travel, leave* (including co-hyponyms *leave, arrive*)
v. meronymy *hand, finger* (including co-meronyms *finger, thumb*)

B: Instantial
i. equivalence The *sailor* was their *daddy*; you'll be the *patient*, I'll be the *doctor*
ii naming The *dog* was called *Toto*; they named the *dog* Fluffy"

Ebenfalls problematisch für die automatische Kohäsionsanalyse ist bei Hasans neuer Klassifikation die Tatsache, dass die neu eingeführten instantiellen Relationen nur durch den Kontext der Äußerung erfasst werden können, der jedoch von Text zu Text unterschiedlich ist. Man kann ohne Schwierigkeit feststellen, dass für die Erkennung dieser Art der lexikalischen Kohäsion komplexes Wissen nötig ist. Deshalb werden in dieser Arbeit nur die lexikalischen Kohäsionsbeziehungen in Betracht gezogen, die unabhängig vom Äußerungskontext erkennbar sind. Neben Wiederholung, Synonymie, Antonymie, Hyponymie und Meronymie, die bei Hasan der allgemeinen lexikalischen Kohäsion zugeordnet sind, werden in dieser Arbeit auch die Ursache-Wirkung-Relation und die Derivation hinzugenommen. Wenn man davon ausgeht, dass sich die kognitiven Funktionen und ihre Merkmale in den lexikalischen Kohäsionsbeziehungen spiegeln können, dann könnte man versuchen, auf dieser Basis weitere Klassifikationen zu unternehmen. Zum Beispiel kann der reduktive Charakter der Wahrnehmung, der in dieser Arbeit im Mittelpunkt steht, als ein Merkmal dienen, nach dem die lexikalische Kohäsion weiter klassifiziert werden kann. Diese Betrachtungsweise ist auch bei Hoffmann (1986:58, 61 ff.) zu finden, der zur Klassifikation der Konzeptmerkmale „Beziehungsmerkmale" eingeführt hat, die sich weiter in vertikale und horizontale Beziehungen einteilen lassen,

wobei die vertikalen Beziehungen der Taxonomie und die horizontalen der Kollokation entsprechen. Auf ähnliche Weise hat Dietze (1994:23 f.) die Beziehungstypen Hierarchie, Assoziationen und Äquivalenz bzw. Identität zur Klassifizierung benutzt. Relationstypen wie Hierarchie und Äquivalenz bzw. Identität lassen sich mit den lexikalischen Relationen Hyponymie und Meronymie sowie Synonymie parallelisieren, während sich der Beziehungstyp der Assoziation kaum durch einfache lexikalische Relationen darstellen läßt.

2.3.3 Lexikalische Kohäsion und Textproduktion

Wissen ist nicht nur das, was in Texten repräsentiert ist, sondern ist gleichzeitig Voraussetzung für die Formulierung von Texten. Wenn Textproduzenten bei der Textproduktion auf einen Sachverhalt Bezug nehmen, mit anderen Worten, wenn ein Text minimalen informativen Charakter zeigt, so ist die entsprechende Wissensstruktur aktiviert, die dem Textproduzenten zur Verfügung steht. Daher ist das Wissenssystem in fast allen Modellen der Sprachproduktion integriert: Zum Beispiel wirkt das Wissen im Schreibmodell von Hayes/Flower (1980) auf die Phase der „Planung“ ein, die mit „Übersetzung“ und „Überprüfung“ eine der drei Hauptkomponenten des Schreibvorgangs ausmacht. Die Auswahl der Wissenselemente bei der Textproduktion bildet im Produktionsmodell von Herrmann (2003) die erste Stufe „Erzeugung der kognitiven Äußerungsbasis“ und ist in drei Teilprozesse untergliedert, nämlich „Fokussierung“, „Parameterfixierung von Teilsystemen der Sprachproduktion“ und „Formatierung“. Zur Fokussierung gehören die sog. „Selektion“ und „Linearisierung“, die für die Auswahl der kognitiven Inhalte und für die Anordnung der Wissenselemente in sequenzieller Reihenfolge, die sich später durch sprachliche Ausdrücke manifestieren, zuständig sind. Bei der „Parameterfixierung von Teilsystemen der Sprachproduktion“ werden situationsspezifische Faktoren wie Lautstärke, Sprachebene, Sie- oder Du-Form etc. eingestellt. Schließlich wird die Botschaft bei der Formatierung in ein Repräsentationsformat umgewandelt. Dieses stellt den Input für die einzelsprachliche Enkodierung dar, bei der das mentale Lexikon eine entscheidende Rolle spielt und bei der die Botschaft in Phonemsequenzen umgewandelt wird. Der phonologische Output stellt schließlich die Basis für die motorische Artikulation dar. Ein ähnliches,

aber umfangreicheres Produktionsmodell ist bei Levelt (1989) zu finden, der das gesamte Produktionssystem in die drei Module „Konzeptualisierung", „Formulierung" und „Artikulation" einteilt. In der Phase der Konzeptualisierung werden die Wissenselemente bzw. Informationen bestimmt, die der Sprecher mitteilen möchte. Die in dieser Phase entstandenen Propositionen[33] werden zum Formulierungsmodul weitergeleitet, wo die passenden Wörter selektiert werden, die syntaktische Struktur bestimmt und die phonologische Enkodierung durchgeführt wird, die die Eingabe für das Artikulationsmodul darstellt, wo schließlich eine für einen Hörer wahrnehmbare sprachliche Ausgabe erzeugt wird.

Kritisch bei der Bestimmung der Wissenselemente für die Textproduktion ist, dass Wissen, das sich eigentlich der konzeptuellen Ebene zuweisen lässt, auch von anderen kognitiven Ebenen, zum Beispiel der referenziellen, aktiviert werden kann. Dieses Problem beruht auf der Ausdifferenzierung der kognitiven Ebenen (vgl. Abschnitt 2.2.2), auf denen die Informationen bzw. das Wissen kontinuierlich verarbeitet werden, wobei die Informationen einer Ebene auch von anderen Ebenen ständig aufgegriffen werden (vgl. Schnotz 1994:78 f.). Zum Beispiel kann die Aktivierung des Wissens auf der konzeptuellen Ebene von Informationen auf der referenziellen Ebene beeinflusst werden und umgekehrt. Dabei lässt sich die genaue Funktionsweise dieser Wechselwirkung schwer erklären. Das Produktionsmodell von Levelt geht von der konzeptuellen Ebene aus, wo das zu übermittelnde Wissen durch „Makroplanung" und „Mikroplanung" bestimmt wird (vgl. Levelt 1989:11).[34] Bei der Makroplanung wird erst die auszudrückende Wissensstruktur bestimmt, diese wird dann weiter in Haupt- und Nebenstruktur gegliedert, und es wird die Reihenfolge der Propositionen (Linearisierung) festgelegt (vgl. Levelt 1989:11, 123-144). Wenn der Sprecher etwa das Konzept UNIVERSITÄT ausdrücken will, so werden diejenigen Teile der konzeptuellen Struktur aktiviert, die mit dem Konzept UNIVERSITÄT in Verbindung stehen. Dazu können FAKULTÄT,

[33] Empirisch ist auch bewiesen, dass Propositionen auf der mentalen Ebene als Bedeutungseinheiten dienen, die für die weitere Verarbeitung in Anspruch genommen werden (vgl. z. B. Christmann 1989: 57, Engelkamp 1976:36-37).

[34] Levelt (1989:72 ff.) berücksichtigt auch die Informationen, die sinneskanalspezifischen Charakter haben und die für die Versprachlichung in die konzeptuelle Ebene überführt werden müssen.

VERWALTUNG, BIBLIOTHEK und sogar KNEIPE gehören, sofern es in der Nähe der Universität viele Kneipen gibt. Wenn das Konzept HERSTELLUNG EINES RECHNERS das Hauptthema ist, so sind wahrscheinlich die Konzepte aktiviert, die den einzelnen Subprozessen des gesamten Herstellungsprozesses entsprechen. Dieser Aspekt ist auch vereinbar mit dem „Prinzip der Konnektivität“ von Levelt (1989:140), das besagt, dass die Konzepte, die mit dem fokussierten Konzept unmittelbar verbunden sind, bei der Auswahl für die Linearisierung gegenüber anderen Konzepten bevorzugt werden. Lexikalische Relationen können als Kanten der Wissensstruktur und auch als Prädikate von Konzepten betrachtet werden. Also entsteht lexikalische Kohäsion, indem die Teile der konzeptuellen Verbindungen durch lexikalische Relationen ausgedrückt werden, die die Wissensstruktur bzw. konzeptuelle Struktur zusammenhalten. Typische Fälle sind Hyperonymie/Hyponymie, Meronymie/Holonymie und kausale Relationen. Die Mikroplanung bestimmt die Präsentation der Informationen, zum Beispiel als Topik oder als Fokus (vgl. Levelt 1989:144-157). Die Mikroplanung trägt pragmatischen Charakter, da die Aktivierung von bestimmten Teilen der Wissensstruktur auf Strategien des Sprechers beruht, die von Kommunikationssituation zu Kommunikationssituation variieren können. Wenn etwa die Kommunikationsteilnehmer die Bibliothek als die wichtigste Universitätseinrichtung ansehen, dann richtet sich ihre Aufmerksamkeit oft auf sie, sodass auch das Konzept BIBLIOTHEK oft aktiviert und schließlich mehrmals erwähnt wird.[35] Ein Konzept kann durch Pronomina, durch Wiederholungen oder durch Hyperonyme und Synonyme wieder aufgenommen werden. Anders als Levelt untersuchen Klein/Stutterheim (1987, 1991) die Informationen auf der referenziellen Ebene. Nach ihnen ist die referenzielle Bewegung, die „die Art und Weise, wie sich die Information innerhalb der […] Referenzbereiche zwischen aufeinander folgenden Äußerungen entwickelt“ (Klein/Stutterheim 1987:166), darstellt, nicht zufällig, sondern beruht auf der Wechselwirkung des „Fraglichen“ („Quaestio“) und der Antworten:

> „Nun geht nicht jeder deklarativen Äußerung eine explizite Frage voraus. Man kann sich aber allemal eine implizite Frage hinzudenken – die

[35] Das Konzept BIBLIOTHEK bzw. das entsprechende Wort der Einzelsprache könnte man dann im Rahmen der funktionalen Satzperspektive als das Thema des geäußerten Satzes bestimmen (vgl. z. B. Lötscher 1983:64-68, Scherner 1984:178 f.).

> Quaestio, die von der betreffenden Äußerung beantwortet wird. […] Die Quaestio einer Äußerung kann sich nun aus einer übergeordneten Quaestio ergeben, nämlich jener, die der Text, zu dem die betreffende Äußerung gehört, in seiner Gesamtheit zu beantworten sucht. Man muss daher zwischen der Quaestio des Textes (der ‚Textfrage') und der einer einzelnen Äußerung unterscheiden." (Klein/Stutterheim 1987:165)

Obwohl man durch die Einbeziehung der referenziellen Bewegung in Texten noch grundlegender auf die Entstehung von lexikalischer Kohäsion eingehen könnte,[36] ist dies für alle kohäsionsbasierten computerlinguistischen Anwendungen nicht immer helfreich. Das lässt sich mit dem folgenden Text demonstrieren:

> „Many wild **bears** have become ‚garbage junkies', feeding from dumps around human developments.To avoid potentially dangerous clashes between them and humans**, scientists** are trying to rehabilitate the **animals** by drugging them and releasing them in uninhabited areas. Although some **biologists** deny that the mind-altering drug was responsible for uncharacteristic behaviour of this particular bears, no research has been done into the effects of giving grizzly **bears** or other mammals repeated doses of phencyclidine." (Hoey 1991:40)

Während *animal* und *bear* im Text denselben Referenten haben, ist dies bei *scientist* und *biologist* nicht der Fall. Obwohl *scientist* und *biologist* nicht koreferent sind, wirken sie bei der Erstellung einer bestimmten konzeptuellen Domäne mit, die zusammen mit der referenziellen Ebene den Kontext des Textes ausmacht.

2.3.4 Lexikalische Kohäsion und Textverstehen

2.3.4.1 Kognitive Verarbeitungsprozesse beim Textverstehen

Textverstehen ist ein äußerst komplexer Vorgang, was bei dessen Untersuchung die Einbeziehung verschiedener Teildisziplinen nötig macht. Die in der

[36] Auf den Zusammenhang von Fragen und lexikalischer Kohäsion hat auch Hellwig (1984:60) hingewiesen: „Fragen und Antworten referieren in bestimmter Weise auf dasselbe im Objektbereich. Zwischen Fragesätzen und Antwortsätzen besteht immer Kohäsion." Man kann weiter einen Zusammenhang von Makro-/Mikroplanung und Quaestio mit den Schemata erkennen, wenn man bedenkt, dass Schemata einen groben Rahmen darstellen, in dem das vorhandene Wissen aktiviert ist und anhand dessen situationsangemessene Informationen erwartet werden. Von daher meint Minsky (1975:246), dass ein Schema eigentlich eine Sammlung von Fragen sei.

linguistischen Pragmatik behandelten Forschungsaspekte[37] können sicherlich zum Teil auch auf das Textverstehen angewandt werden, aber die sich daraus ergebenden Textanalysemodelle sind nicht genügend auf den Textrezipienten und damit auf die konstruktiven Aspekte bezogen, die den Aufbau von Wissen erklären. Erst der Forschungsansatz der Kognitionswissenschaft ermöglicht bei der Textverstehensforschung die angemessene Berücksichtigung der kognitiven Mechanismen, die für die Textrezeption notwendig sind.[38] Beim Textverstehen werden im Unterschied zur direkten Wahrnehmung die Informationen nicht direkt, sondern durch sprachliche Zeichen zum Textrezipienten übertragen, der indirekt ein (mögliches) Wirklichkeitsmodell konstruiert (vgl. Engelkamp 1984a:32 f.). Givón (1995:64) meint, dass „text comprehension is synonymous with the construction of a structured mental representation of the text". Aufgrund dieses konstruktiven Aspekts ist es nicht besonders verwunderlich, dass zahlreiche Textverstehensmodelle von der Schematheorie und ihren Annahmen über den Erwerb und Aufbau des Wissens geprägt sind. Besonders bedeutsam ist das Textmodell von van Dijk (1980), das im nächsten Kapitel vorgestellt wird (vgl. aber auch Ballstaedt/Madle/Schnotz 1981:41-84, Mandl/Friedrich/Hron 1988, Minsky 1975, Rumelhart 1975, Thorndyke 1977). Textverstehen kann als ein mentaler Prozess aufgefasst werden, der von der Erkennung des ersten Buchstabens eines Textes bis hin zur Speicherung des Resultats des Verstehens zeitlich sequenziell abläuft. Die Analyse dieses Prozesses ist nicht einfach, da beim Textverstehen sehr viele unterschiedliche kognitive Teilprozesse, die sprachbezogen sind, parallel und kontinuierlich ablaufen. Trotz dieser Schwierigkeiten gelangt Scherner (1989:94-97) zur Unterscheidung von vier Arten von Prozessen, die für das Textverstehen relevant sind: 1. Subsemantische Verarbeitungsprozesse, 2. semantisch-syntaktische Verarbeitungsprozesse, 3. elaborative Verarbeitungsprozesse, 4. reduktive Verarbeitungsprozesse. Subsemantische Verarbeitungsprozesse sind für die Worterkennung zuständig; die semantisch-syntaktischen Verarbeitungsprozesse filtern den Inhalt der Sätze heraus, die elaborativen Verarbeitungsprozesse ermöglichen Schlussfolgerungen auf das nicht explizit ausgedrückte Wissen.

[37] Zur kommunikationsorientierten Textlinguistik vgl. z. B. Brinker 2005:15 f., Hartung 2000.
[38] Christmann (2003:113) definiert z. B. Textverstehen als „Interaktion zwischen einem vorgegebenen Text und der Kognitionsstruktur des/der Rezipienten/in."

Dieser Verarbeitungsschritt dient der Ermittlung von Propositionen,[39] die im Text zwar nicht explizit durch Sätze repräsentiert sind, die aber für die Textkohärenz und damit für das Textverstehen notwendig sind. Durch die reduktiven Verarbeitungsprozesse, die für die vorliegende Arbeit von großem Belang sind, werden die aus den anderen Prozessen entstandenen Propositionen in eine oder mehrere Propositionen kondensiert, die den ganzen Textinhalt zusammenfassend darstellen. Darüber wird im nächsten Abschnitt ausführlicher gesprochen, und es wird gezeigt, dass Reduktion auch auf die Funktion von Schemata zurückzuführen ist.

2.3.4.2 Informationsreduktion beim Aufbau von Textrepräsentationen: Das Textmodell von van Dijk (1980)

Im Textmodell von van Dijk (1980) kommt den Propositionen und ihrer Konnexion in der „Makrostruktur“ und „Mikrostruktur“ eine besondere Bedeutung zu. Die Propositionen der Sätze im Text bilden durch ihre lineare Konnexion Mikrostrukturen, aus denen durch einen Bottom-up-Prozess die darüber liegende Makrostruktur gebildet wird. Die Makrostruktur ist eine abstrakte semantische Struktur und beruht auf der Annahme, dass ein Text als Ganzes einen „globalen Zusammenhang“ besitzt (vgl. van Dijk 1980:39-43). Sie ist so entscheidend für van Dijks Textmodell, dass er nur solche Satzsequenzen als Texte ansieht, die eine Makrostruktur besitzen.[40] Das allgemeine Schema für Makrostrukturen ist in Abbildung 2-7 dargestellt. Jeder Knoten in der Struktur repräsentiert eine Proposition. Die Propositionen einer Ebene bilden durch ihre lineare Konnexion eine Mikrostruktur. Die Makrostruktur ist von hierarchischer Art und entsteht dadurch, dass die Propositionen der jeweils darunter liegenden Ebene zu einer Proposition zusammengeführt sind, die diese Propositionen inhaltlich zusammenfasst.

[39] Ein vollständiges Textverstehen beansprucht nicht nur Propositionen, die sich der konzeptuellen Ebene zuweisen lassen, sondern auch den Bezug auf die reale Welt. Dies wird schnell deutlich anhand des Problems der Koreferenz, das sich auf bestimmte kommunikative Situationen bezieht. Das sog. Situationsmodell von van Dijk/Kintsch (1983:336-345) und das mentale Modell von Johnson-Laird (1989) basieren auf dem Grundgedanken, dass ein Textverstehensmodell auch die referenzielle Ebene einbeziehen muss.

[40] Van Dijk (1980:41): „Nur die Satzsequenzen, die eine Makrostruktur besitzen, werden wir (theoretisch) als Texte bezeichnen.“

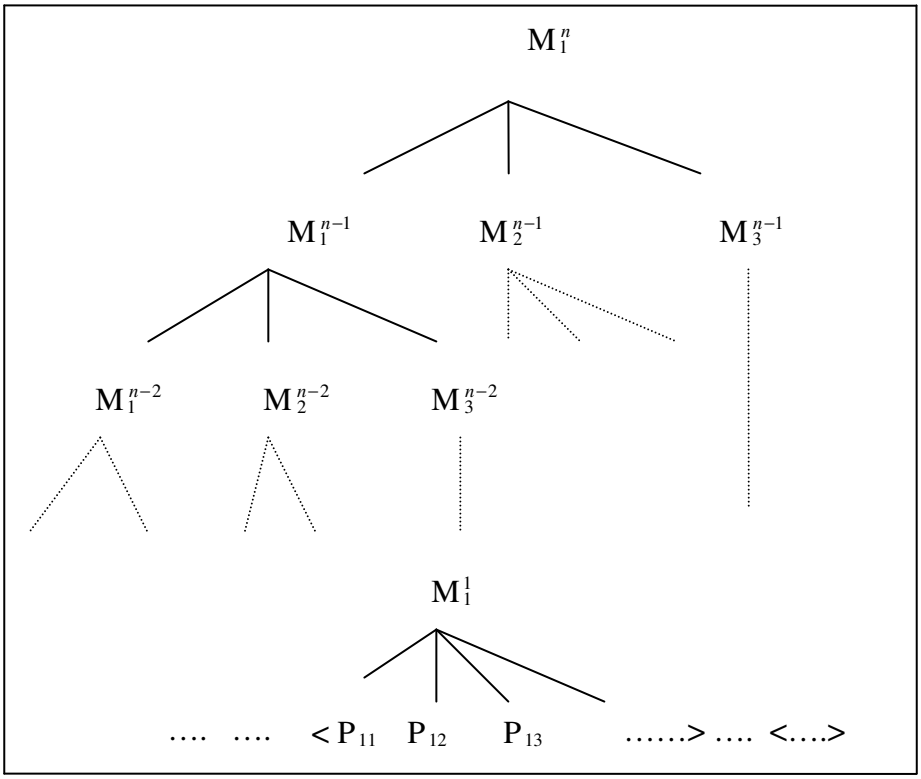


Abbildung 2-7: Allgemeines Schema für Makrostrukturen (van Dijk 1980:43)

Da der Begriff „Makrostruktur“ relativ zur Betrachtungsebene bestimmt ist, kann M_1^{n-1} als eine Proposition betrachtet werden, die zusammen mit M_1^{n-2}, M_2^{n-2} und M_3^{n-2} die Mikroebene von M_1^n bildet und gleichzeitig als eine Proposition auf der Makroebene betrachtet wird, die M_1^{n-2}, M_2^{n-2} und M_3^{n-2} subsumiert. Dementsprechend können verschiedene Makrostrukturen in einem Text bestehen, wobei die auf der höchsten Ebene stehende Struktur, die in Abbildung 2-7 durch M_1^n und die darunter stehenden Ebenen M_1^{n-1}, M_2^{n-1}, und M_3^{n-1} gekennzeichnet wird, als „Makrostruktur des Textes“ gilt, da sie „die allgemeinste und globalste Makrostruktur des Gesamttextes“ darstellt.[41] Die eine Makrostruktur bildenden Relationen lassen sich in zwei Typen klassifizieren, und zwar in die Relationen zwischen Propositionen auf derselben Mikroebe-

[41] Der Begriff „Makrostruktur“ wird von van Dijk (1980:45) auch mit dem Themabegriff in Verbindung gebracht. Demnach sind Makropropositionen Themen von Textteilen (bzw. -ausschnitten), während die auf der obersten Ebene stehende Makroproposition, die alle darunter angeordneten Propositionen zusammenfasst, das Textthema darstellt.

ne[42] und in die hierarchiebildenden Relationen. Für die Bildung von hierarchischen Makrostrukturen aus den Propositionen einer Mikroebene benötigt man semantische Transformationsregeln, die van Dijk (1980:43 ff.) „Makroregeln" nennt und die den Schemafunktionen der Schematheorie entsprechen (vgl. Rickheit/Strohner 1992:80). Van Dijk (1980:45-67) nimmt folgende vier Makroregeln an: 1. Auslassen, 2. Selektieren, 3. Generalisieren, 4. Konstruieren bzw. Integrieren. Nach Regel 1 werden alle Propositionen ausgelassen, deren Informationen für den Gesamtzusammenhang unwichtig, nebensächlich sind. Abbildung 2-8 zeigt die Anwendung von Regel 1:

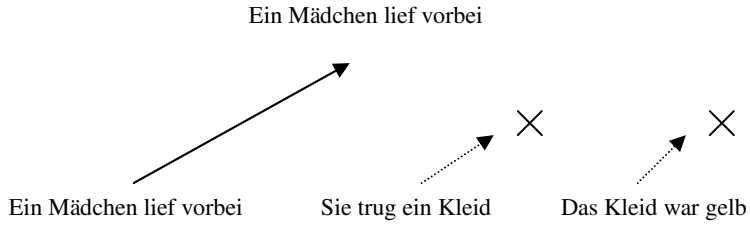


Abbildung 2-8: Auslassen

In die Makroproposition, die durch den Satz *Ein Mädchen lief vorbei* dargestellt ist, geht nur der Inhalt der gleichlautenden Proposition der Mikroebene ein, die anderen Propositionen fallen weg. In einem anderen Kontext wäre es sicherlich auch möglich, dass weitere Propositionen ausgewählt werden, etwa in einem Kontext, in dem die Farbe des Kleids oder die Tatsache, dass das Mädchen überhaupt ein Kleid trug, von Relevanz für den übergeordneten Textsinn sind. Dann könnte die Makroproposition durch die Sätze *Ein Mädchen in einem Kleid lief vorbei.* bzw. *Ein Mädchen in einem gelben Kleid lief vorbei.* ausgedrückt werden.

[42] Einige Relationseigenschaften zwischen Propositionen hat van Dijk (1980:31) wie folgt vorgestellt:
„ (i) A ist Ursache von B (= B ist Folge von A).
(ii) A ist eine Begründung von B (wobei B eine Handlung ist oder die Folge einer Handlung).
(iii) A und B ereignen sich in derselben Situation (d. h.: im Weltzeitpaar <w_i, ti>) und gehören demselben konzeptionellen Bereich an; zugelassen ist:
– A ist gleichzeitig mit B;
– A findet in einer Teilperiode von B statt (oder umgekehrt);
– A und B folgen aufeinander (wie in der kausalen Beziehung);
– A und B überlappen einander.
(iv) A ist notwendigerweise (logisch, konzeptionell) Teil von B, oder umgekehrt.
(v) A ist ein normaler (konventioneller) >Bestandteil< von B, oder umgekehrt."

Nach Regel 2 werden alle redundanten Propositionen gestrichen, das heißt solche, deren Informationsgehalt aus anderen Propositionen erschlossen werden kann; nur die inhaltlich notwendigen Propositionen werden selektiert.

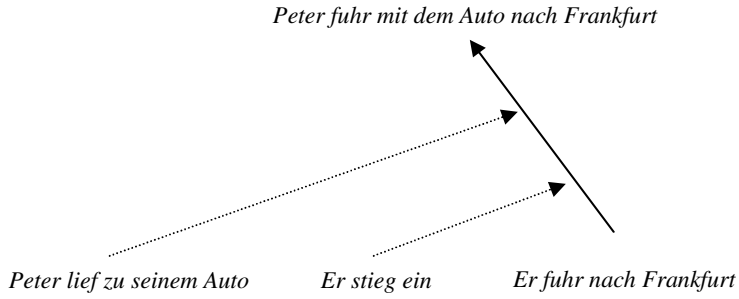


Abbildung 2-9: Selektieren

Abbildung 2-9 zeigt die Anwendung der Regel 2. Die drei Aussagen *Peter lief zu seinem Auto. Er stieg ein. Er fuhr nach Frankfurt.* lassen sich durch das Streichen der redundanten Propositionen (bzw. der Selektion der notwendigen Propositionen) zur Proposition *Peter fuhr mit dem Auto nach Frankfurt.* reduzieren. Wenn Peter mit dem Auto nach Frankfurt fährt, dann ist logisch impliziert, dass er vorher zu seinem Auto lief und in das Auto einstieg. Diese Propositionen sind also redundant und können daher gestrichen werden.
Der Reduktionsprozess ist bei den Regeln 3 (Generalisieren) und 4 (Konstruieren oder Integrieren) noch komplexer, da hier auch neue Konzepte gebildet werden.

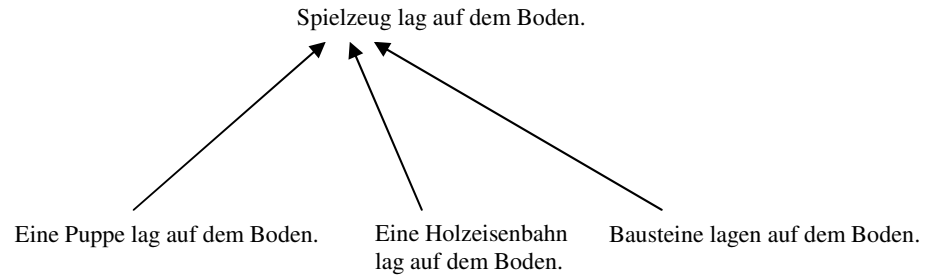


Abbildung 2-10: Generalisieren

Die Anwendung der Regel 3 ist in Abbildung 2-10 demonstriert. Die neue Proposition lautet *Spielzeug lag auf dem Boden*. Das Konzept *Spielzeug* enthält die

gemeinsamen semantischen Merkmale von *Puppe*, *Holzeisenbahn* und *Bausteine*, während spezifischere Merkmale ausgelassen sind, die beispielsweise eine Holzeisenbahn von einer Puppe unterscheiden. Da *Spielzeug* das Hyperonym der Kohyponyme *Puppe*, *Holzeisenbahn* und *Bausteine* darstellt, zeigt dieses Beispiel sehr schön die wichtige Rolle von hierarchischen lexikalischen Relationen, insofern als diese die kognitiven Reduktionsprozesse auf der Textebene repräsentieren. Regel 4 (Konstruieren oder Integrieren) ist für das Textverstehen von großer Relevanz, weil sie für den Reduktionsprozess nicht mehr nur auf die Propositionen zurückgreift, die aus dem Text gewonnen wurden, sondern auf Schemata, die unser stereotypes Wissen über typische Situationen, Ereignisse oder Handlungen und ihre Bedingungen, Folgen, Begleitumstände etc. repräsentieren. Bei der Anwendung von Regel 4 werden Mikropropositionen nicht durch Auslassen oder Generalisieren reduziert, sondern dadurch, dass aus dem aktivierten Frame/Schema neue Informationen erschlossen werden, die das gemeinsame „Thema" der Mikropropositionen enthalten (vgl. van Dijk 1980:48). Abbildung 2-11 zeigt, wie die vier Mikropropositionen *Ich ging zum Bahnhhof. Ich lief zum Bahnsteig. Ich stieg in den Zug ein.* und *Der Zug fuhr ab.* das Thema „Zugreise" bzw. die Proposition MACHEN (ICH, ZUGREISE) implizieren:

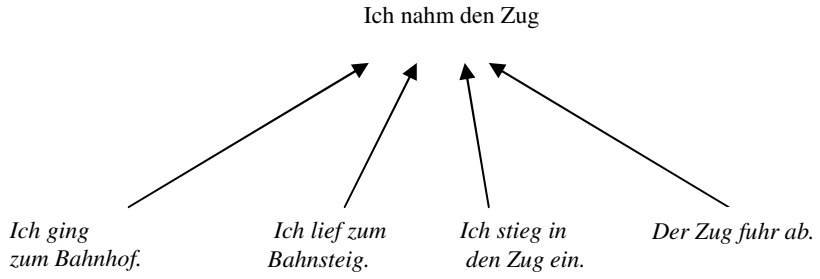


Abbildung 2-11: Konstruieren

2.3.4.3 Lexikalische Relationen als gemeinsames Strukturmerkmal des mentalen Lexikons und der Makrostruktur von Texten

Das Reduktionsmodell von van Dijk ist sicherlich übersimplifiziert. Neben der im Mittelpunkt stehenden inhaltlich-propositionalen Ebene gibt es auch andere

Ebenen wie etwa die „funktional-illokutive Ebene der Handlung“[43] oder die „sprachlich-ausdrucksseitige Ebene der grammatischen Einheiten“, die bei der Textkonstitution zum Tragen kommen (vgl. Nussbaumer 1991:161). Dennoch kommt der Analyse der inhaltlich-propositionalen Ebene die größte Bedeutung für die automatische Indexierung zu, da der deskriptive Aspekt von Texten bei der Indexierung eine zentrale Rolle spielt (vgl. Hjøland 1997:91). Eine Proposition besteht aus einem Prädikat und einem oder mehreren Argumenten (vgl. Engelkamp 1976:21-25). Auch lexikalische Relationen können als zweistellige Prädikate verstanden werden, die zusammen mit ihren Argumenten – dem Wortpaar, das unter der jeweiligen lexikalischen Relation steht – eine Proposition bilden. Das lexikalische Relationsverhältnis zwischen zwei Wörtern X und Y kann deshalb durch Propositionen der Form LEXIKALISCHE RELATION (X, Y) zum Ausdruck gebracht werden, zum Beispiel durch Aussagen wie X *ist ein Synonym von* Y oder X *ist ein Meronym von* Y. Das lexikalische System, das jedem Individuum auf der mentalen Ebene zur Verfügung steht, stellt die Grundausrüstung für den Aufbau neuen Wissens dar, das normalerweise durch Propositionen repräsentiert ist. Diese Arbeit geht von der Prämisse aus, dass das mentale Lexikon eine Wissensrepräsentation darstellt, für deren Aufbau das gleiche kognitive Prinzip relevant ist wie für den Aufbau einer konzeptuellen Textrepräsentation, nämlich die Reduktion und Verdichtung von Information anhand der Schemafunktion. Folgt man dieser Annahme, dann kann das lexikalische System selbst als eine Art Makrostruktur angesehen werden. Die Informationsreduktion zeigt sich an der Existenz von hierarchischen Relationen wie Hyponymie/Hyperonymie und Meronymie/Holonymie sowohl im mentalen Lexikon als auch in der kognitiven Makrostruktur von Texten.

[43] In der Linguistik ist vor allem die pragmatische Ansicht vom Text als sprachlicher Kommunikationshandlung bzw. als Mittel oder Träger für die Intention des Textproduzenten beherrschend (vgl. z. B. Brinker 2005:15 f., Motsch 2000:415).

3. Anwendung auf die Indexierung

3.1 Prinzipien der Indexierung

Dass die Wissensbasis für praktische automatische Indexierung nur begrenzt das menschlichen Wissen ersetzen kann, macht die Veränderung der Betrachtungsweise der drei bekannten Indexierungsprinzipien notwendig (vgl. Cleveland/Cleveland 2001:98, Lancaster 1998:5-19): „aboutness“, „exhaustivity“ und „specificity“. Die „aboutness“ bezieht sich auf der Frage, worum es sich bei dem betreffenden Dokument handelt. Diese Frage ist sogar bei der traditionellen manuellen Indexierung sehr kritisch, da es verschiedene Aspekte gibt, die bei der Inhaltsanalyse berücksichtigt werden können, und die „aboutness“ der Informationen je nach den Voraussetzungen der Benutzer unterschiedlich sein kann (vgl. Cleveland/Cleveland 2001:98, Hjøland 1997:61-67, Lancaster 1996:10-14). Von daher ist es nicht besonders verwunderlich, dass für die automatische Indexierung der Begriff „aboutness“ sehr allgemein aufgefasst wird. Also sollen Indexterme möglichst wichtige inhaltliche Aspekte darstellen, wenn sie dem Zweck des Retrievals dienen. Es ist in der automatischen Indexierung üblich, dass Nomen repräsentativ für den Textinhalt gelten und daher als Indexterme verwendet werden (vgl. Baeza-Yates/Ribero-Neto 1999:163). „Exhaustivity“ und „specificity“ sind zwei Prinzipien, die sich auf die zwei verschiedenen Aspekte der Granularität der Indexierung beziehen. Ersteres besagt, dass möglichst viele Indexterme berücksichtigt werden sollen, damit mehrere Zugangsmöglichkeiten zu Dokumenten zur Verfügung stehen. Das große Verdienst der automatischen Indexierung bezieht sich auf diesen quantitativen Aspekt der Indexierungsprinzipien. „Specificity“ wiederum besagt, dass Indexterme für die Darstellung der Dokumente möglichst spezifisch sein sollen. Dies entspricht dem qualitativen Aspekt der Indexterme, der üblicherweise „Präzision“ genannt wird. Dieser Aspekt wird durch die Angabe der Gewichtung der Indexterme in der automatischen Indexierung realisiert.

3.2 Indexsystem und Komponenten des Information Retrieval System (= IRS)

Da das Endziel einer Indexierung immer Retrieval ist (vgl. Kowalski 1997:48), ist es undenkbar, Indexierung unabhängig von IRS zu berücksichtigen.

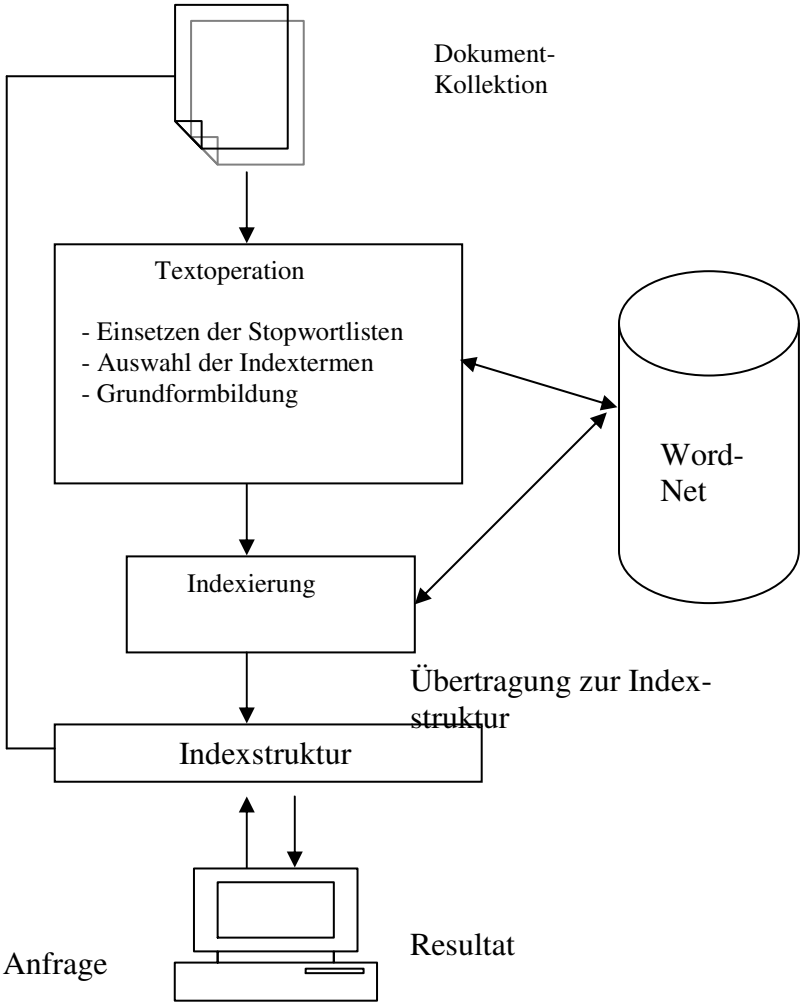


Abbildung 3-1: Komponenten des IRS

Diese enge Verbindung verursacht besonders in der automatischen Indexierung manchmal die Schwierigkeit, zwischen Indexierungs- und Retrievalprozessen eine klare Trennlinie zu ziehen.[44] Dies zeigt einerseits den Status des Indexsystems im IRS und andererseits die weiteren Zusammenhänge zwischen Indexie-

[44] Zum Beispiel wird das Vektorraummodell sowohl als Indexierungs- wie als Retrievalmodell betrachtet (vlg. Baeza-Yates/Ribeiro-Neto1999:24-30, Stock 2007:191-206) .

rungsprozess und anderen Komponenten im IRS. Wie Abbildung 3-1 darstellt, sind die grundlegenden Prozesse des IRS, die für diese Arbeit entworfen wurden, nicht sehr unterschiedlich von den gängigen Modellen des IRS.[45] Das IRS besteht in dieser Arbeit aus vier Subprozessen: Textoperation, Indexierungshauptprozess, Übertragung in die Indexdatenbank und Eingabe der Anfrage.

Durch den engen Zusammenhang des Indexierungsprozesses mit IRS wird ersichtlich, dass der Textoperationsprozess nach der Indexierungsstrategie bestimmt werden muss und dass die Anfragesprache auch mit der Repräsentationsform des Indexterms übereinstimmen muss. In diesem Abschnitt werden einige Aspekte des ersten, dritten und vierten Prozesses behandelt, auf den zweiten Subprozess wird im nächsten Abschnitt eingegangen.

3.2.1. Textoperation

Die Textoperation lässt sich in drei Subprozesse einteilen: Einsetzen der Stoppwortlisten, Auswahl der Indexterme und Grundformbildung.

Stoppwortliste

Im Information Retrieval (IR) ist es üblich, mit Hilfe einer Stoppwortliste zunächst diejenigen Wörter, die semantisch von geringer Bedeutung sind, herauszufiltern, wodurch die Zahl der zu analysierenden Wörter reduziert wird.[46] Der Einsatz einer Stoppwortliste ist im Rahmen der linguistischen Textanalyse – insbesondere wenn lexikalische Ketten betrachtet werden – nicht ganz unproblematisch, da die weitere Fortführung eines Konzepts im Text nicht nur durch die Wiederholung der Wörter, sondern auch durch Hyponyme wie *thing* oder *man* sprachlich dargestellt wird, die manchmal als semantisch unwichtig gelten und sich daher in manchen Stoppwortlisten finden. Dies hängt mit dem Problem der Bestimmung des Transitivitätsgrades bei der Erkennung der Hyponymie zusammen. Wie lang darf also die Pfadlänge im hyponymen Baum bei der Erkennung von Hyponymie werden? Auf dieses Problem wird später noch einmal eingegangen. Insbesondere bei der gezielten Suche, wenn nur eine spe-

[45] Zu den Prozessmodellen des IRS s. Baeza-Yates/Ribeiro-Neto 1999:9 f., Lancaster 1997:1-4.
[46] Baeza-Yates/Ribero-Neto (1999:167) weisen darauf hin, dass die Zahl der Indexterme durch eine Stoppwortliste auf bis zu 40 % reduziert werden kann.

zifische Domäne von Interesse ist, wäre es zwar wünschenswert, dass die Stoppwortliste dem Zweck und der Graduierung der Suche angepasst ist. Da das in dieser Arbeit nicht der Fall ist, wird die Stoppwortliste von Fox (1992) eingesetzt, die ungeachtet einer bestimmten Domäne hergestellt wurde.

Auswahl der Indexterme

Es ist üblich im IR, dass Nomen als Indexterme selektiert werden (vgl. Baeza-Yates/Ribero-Neto 1999:169). Dies trifft auch für die meisten kohäsionsbasierten Textanalysen zu (vgl. z. B. Al-Halimi//Kazman 1998, Stairmand 1997, Vohees 1998). Es ist sicherlich nicht zu verneinen, dass eine semantische Ähnlichkeit zwischen den Wörtern, die nicht zu einer Wortklasse gehören, bestehen kann, und es ist zu erwarten, dass durch die Berücksichtigung anderer Wortklassen wie Verben und Adjektive eine bessere Ermittlung der semantischen Informationen erreicht wird. In diesem Fall ist in Kauf zu nehmen, dass es eine erhebliche Erhöhung der Komplexität der Algorithmen bei der Erkennung von lexikalischen Verbindungen gibt. In dieser Arbeit wird versucht, neben Nomen auch Verben zu berücksichtigen, wobei die semantische Verbindung von Wörtern mit Nomen und Verben durch Derivation ermittelt wird, die nicht als lexikalische Relation im strengen Sinne gilt.

Erkennung der Grundformen

Weitere Analyseeinheiten sind gemäß den jeweiligen Verfahrensweisen der Indexierung unterschiedlich bestimmt. Für die Bestimmung der Analyseeinheit werden „Stemming“ und „Grundformbildung“ üblicherweise im IR verwendet, wobei Letzteres normalerweise als „Lemmatisierung“ bezeichnet wird (vgl. Stock 2007:228-247). Durch das sogenannte Stemming wird die Stammform eines Wortes erkannt,[47] wobei Informationen, die mit dem Präfix bzw. Suffix verbunden sind und eventuell für die Erkennung der Wortklasse verwendet werden können, verlorengehen. Da die Erkennung der lexikalischen Relationen einen der wichtigsten Teilprozesse in dieser Arbeit darstellt, ist die Bestimmung der grundlegenden Analyseeinheiten, auf denen die lexikalischen Relationen basieren, eine wichtige Voraussetzung. Wie in Abschnitt 2.1 gezeigt, setzen lexikalische Relationen voraus, dass zwei Wörter lexikalische Einheiten

[47] Einige Beispiele des Stemmingverfahrens s. Moens 2000:81 ff.

sind und diese zur einer Wortklasse gehören. Um ein Wort als eine lexikalische Einheit zu erkennen, ist es nötig, seine Grundform zu ermitteln. Zwei Verfahren zur Erkennung von Grundformen sind vorzustellen: Durch das sog. „regelgeleitete Verfahren“ wird die Grundform ohne Lexikon hauptsächlich anhand morphologischer Regeln erstellt.[48] Hingegen wird das „wörterbuchbasierte Verfahren“ dadurch gekennzeichnet, dass Grundformen durch den Einsatz eines Lexikons erkannt werden (vgl. Nohr 2001:56-60, Stock 2007:229 ff.). Lemmata bzw. lexikalische Einheiten lassen sich durch ihre Grundform darstellen und schließen normalerweise die Information über ihre Wortklasse ein (vgl. Abschnitt 1.3.4). Problematisch ist, dass einer Form mehrere Wortklassen zugewiesen werden können, zum Beispiel *convert* als Verb und Nomen und *converts* als jeweils durch Konjugation oder Plural realisierte Form. Dieses Problem ist besonders bei Hausser gut berücksichtigt, der den ganzen Worterkennungsprozess in Kategorisierung und Lemmatisierung einteilt, wobei er die Information über die Wortklasse in der „Kategorisierungsphase“ ermittelt und damit weiter die Erkennung der Grundform durchführt (vgl. Hausser 1998:38-57). In dieser Arbeit wird die Grundform der Wörter anhand der von WordNet bereitgestellten Funktionen ermittelt, die als Eingabeparameter eine der möglichen Variationen der Grundform und die Angabe über die Wortklasse benötigt, die beide aus dem annotierten Teil des Brownkorpus entnommen sind.

Korpus

In dieser Arbeit wird als Testdokument das Brownkorpus (Francis/Kucera 1979) verwendet, das 500 Dokumente und insgesamt 1014312 Wörter enthält, wobei jedes Dokument aus ca. 2000 Wörtern besteht. Von den 1014312 Wörtern des Brownkorpus haben die IR-Systeme der beiden verschiedenen Ansätze anhand der Stoppwortliste und der Auswahl von Nomen und Verben 657313 Wörter analysiert und 14565 Indexterme bestimmt.

Das Brownkorpus besteht aus verschiedenen Textsorten, die sich vor allem in informative und fiktive Texte differenzieren. Es besteht aus 374 informativen

[48] Als Beispiel für ein regelgeleitetes Verfahren für Englisch ist der „S-Lemmatiser“ zu nennen, über dessen Vorgang Stock (2007:229) folglich zusammenfasst: „1. Wenn eine Wortform drei Buchstaben oder weniger umfasst, beende die Bearbeitung. 2. Wenn eine Wortform mit IES, aber nicht mit EIES oder AIES endet, dann ersetze IES durch Y. 3. Wenn eine Wortform mit ES, aber nicht mit AES, EES oder OES endet, dann ersetze ES durch E. 4. Wenn eine Wortform mit S, aber nicht mit US oder SS endet, dann lösche das S.“

und 126 fiktiven Texten, die in dieser Arbeit für den Ausgabevergleich von 1 bis 500 nummeriert sind. Die Dokumente mit Doc-Nr. 1 bis 374 sind also als informative, die mit Doc-Nr. 375 bis 500 als fiktive Texte kategorisiert. Besonders wichtig für den Vergleich sind 116 Texte: die Texte mit der Doc-Nr. 106 bis 141, die von Hobbys und Fertigkeiten handeln, beispielsweise ein Kochrezept (Doc-Nr. 119) oder eine Beschreibung der Gartenarbeit (Doc-Nr. 107), und die Texte mit der Doc-Nr. 295 bis 374, die als Lernmaterial für verschiedene wissenschaftliche Domänen geschrieben sind.

3.2.2 Indexstruktur

Zwei Möglichkeiten sind vorstellbar beim Suchen der Anfragewörter in der Dokumentkollektion (vgl. Baeza-Yates/Ribeiro-Neto 1999:191-228, Stock 2007:132-127): Erstens: Die Anfragewörter werden in allen Dokumenten der Dokumentkollektion sequenziell gesucht. Diese Methode ist nicht sehr effizient, wenn die Dokumentkollektion groß ist. Die zweite Möglichkeit, die sich als bessere Alternative erweist, ist die Verwendung einer sogenannten invertierten Datei, die die Indexterme mit den dazugehörigen Informationen wie Referenz zu den Dokumenten und den Stellen des Dokuments, in denen die Wörter vorkommen, enthält. In dieser Arbeit wird eine invertierte Datei verwendet, wobei deren Implementierung die zeichenbasierte Tries-Datenstruktur zugrunde liegt (Implementation nach Thaller 1989). Abbildung 3-2 stellt dar, wie die vier Indexterme *apple*, *appeal*, *digit* und *digital* in der Indexstruktur verwaltet werden.

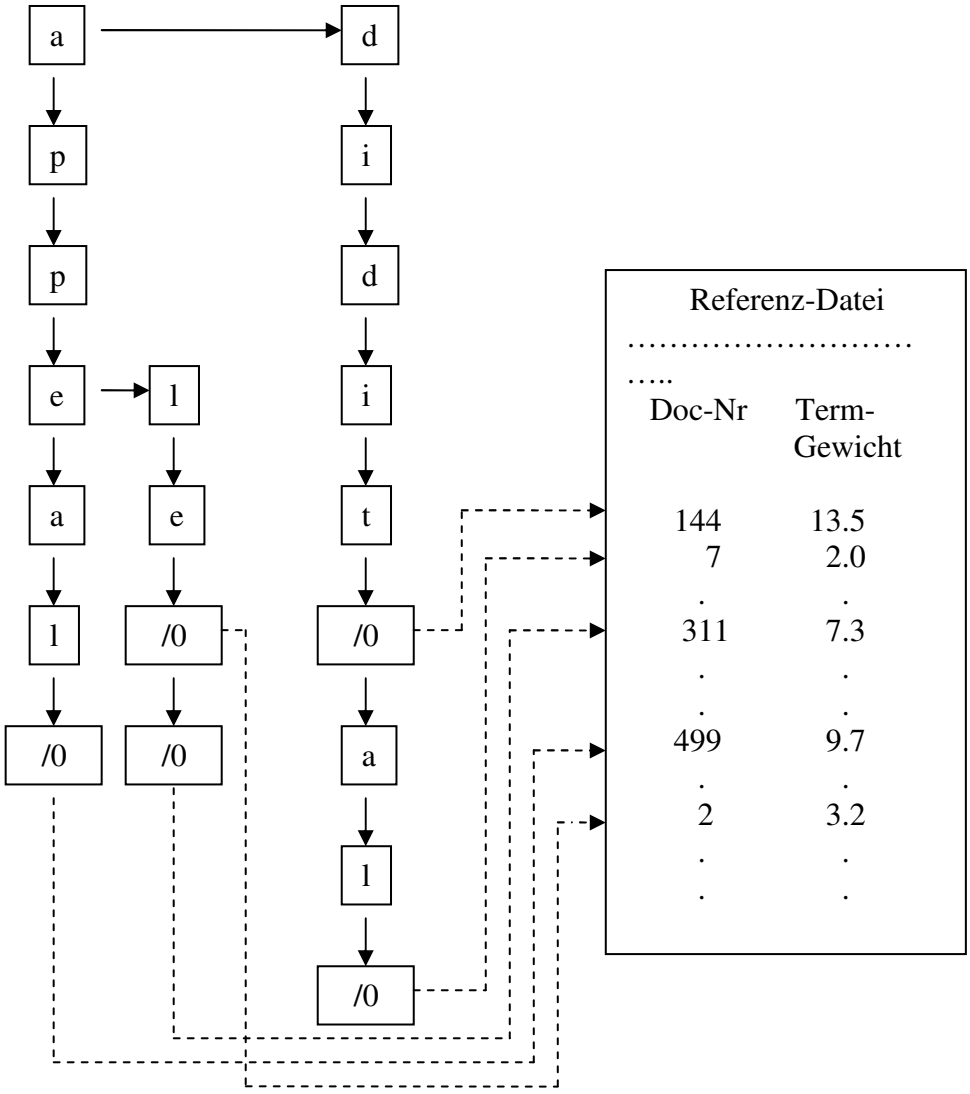


Abbildung 3-2: Tries-basierte Indexstruktur (leicht geändert nach Thaller 1989:5)

Knoten entsprechen einzelnen Zeichen, aus denen ein Indexterm bestehen kann. Die Knoten sind durch Zeiger verbunden, die hier durch Pfeile dargestellt sind, wobei logisch gesehen $\rightarrow$ für ODER und $\downarrow$ für UND steht. Speicherplatz wird vor allem dadurch gespart, dass gemeinsame Zeichen der Indexterme nur einmal gespeichert werden, zum Beispiel „app“ von „apple“ und appeal“ und „digit“ von „digit“ und „digital“. Ein Indexterm ist mit „/0“ terminiert, und gleichzeitig enthält die Referenz-Datei die Informationen über das Vorkommen des Indexterms in der Dokumentkollektion.

3.2.3 Die Granularität von Wissensrepräsentation und Anfrage

Was Semantik im IRS angeht, gibt es mindestens drei kognitive Subjekte, die ein eigenes semantisches Wissen besitzen: Dokumentverfasser, Benutzer und Wissensbasis, in diesem Fall WordNet. Wenn die Indexierungsprozesse nach einer Wissensbasis durchgeführt werden, dann heißt dies auch, dass das daraus sich ergebende Indexsystem unter dem Einfluss der Wissensstruktur der Wissensbasis steht, die sich wiederum von der Wissensstruktur des Benutzers unterscheiden kann. Zum Beispiel bietet WordNet die folgenden sieben Interpretationsmöglichkeiten zu dem Wort *school*, wenn dieses als Nomen verwendet wird:

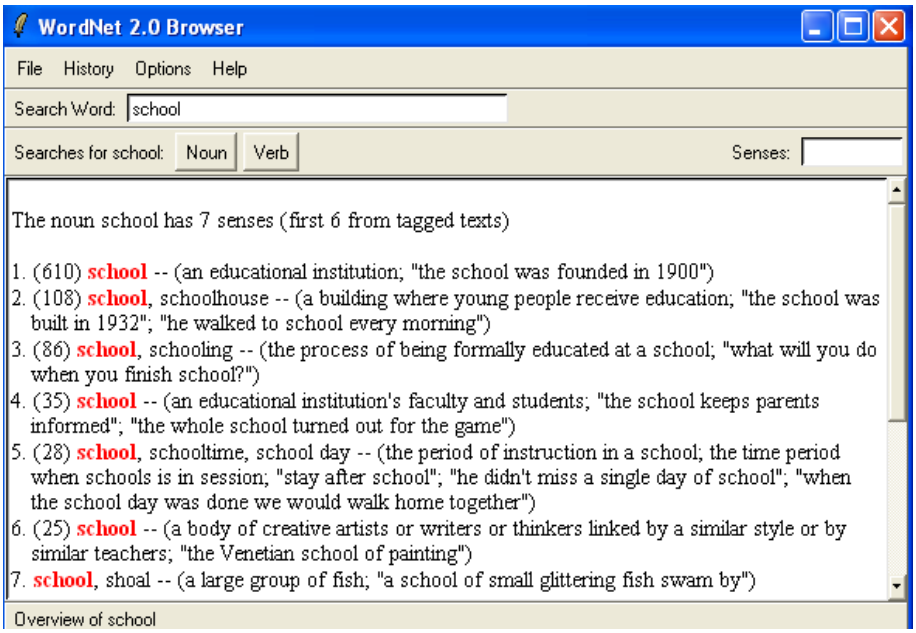


Abbildung 3-3: Mögliche Belegformen von „school“

Wenn alle Bedeutungsdifferenzierungen des Benutzers von *school* relevant wären, dann wären auch sieben getrennte Indexterme nötig, die jeweils der einzelnen Bedeutung bzw. dem einzelnen Konzept entsprechen. Weiter soll dem Benutzer die Möglichkeit geboten werden, solche verschiedenen Konzepte mit der Anfragesprache repräsentieren zu können. An dieser Stelle sind zwei Faktoren zu erkennen, die bei dem Einsetzen einer Wissensbasis bei IRS in Betracht gezogen werden sollen: Erstens soll die Granularität der Wissensrepräsentation nach Interesse des Benutzers und dem von ihm benötigten Wissen

bestimmt werden. Zweitens sollen die Granularitäten der Wissensrepräsentation und der Repräsentation der Informationsbedürfnisse, die normalerweise durch die Anfragesprache dargestellt werden, möglichst einheitlich sein. In dieser Arbeit, die nicht domänenspezifisches Suchen, sondern allgemeines Suchen vornimmt, wird keine Spezifizierung und Erweiterung der Anfragesprache vorgenommen. Aber weil das Wissen in WordNet für das allgemeine Suchen zu spezifisch bestimmt ist, empfehlt es sich, dass der Indexierungsprozess darauf Rücksicht nimmt, diese hohe Spezifizierung möglichst zu mindern. Dies wird bei der Berechnung der Indextermgewichte berücksichtigt, die in Abschnitt 3.3.5 behandelt wird.

3.3 Anwendung der kognitiven Interpretation von lexikalischer Kohäsion in der Indexierung

3.3.1 Probleme bei der Erkennung der lexikalischen Kette

Im Text kann ein einzelnes Wort mit mehreren anderen Wörtern durch verschiedene lexikalische Relationen in Verbindung stehen, und diese Wörter können wieder mit anderen Wörtern verbunden sein. Damit entstehen mindestens folgende drei Arten von Informationen:

i) die möglichen Belegformen zu Wortformen, auf denen lexikalische bzw. semantische Relationen beruhen;
ii) die entsprechende Wortform;
iii) die lexikalischen Relationen, durch die die Konzepte verbunden sind.

Durch die Informationen bilden sich sog. lexikalische Ketten, die als eine Art Cluster bzw. Klasse betrachtet werden können. Wenn beispielsweise die Wörter *apple, fruit*, *banana* und *crabapple* in dieser Reihenfolge im einem Text vorkommen und die lexikalische Relationen zwischen *fruit*:*apple*, *apple*:*crabapple* und *fruit*:*banana* erkannt sind, dann entsteht eine lexikalische Kette, die so wie in Abbildung 3-4 aussehen kann, wobei Hyperonymie/Hyponymie durch Pfeile gekennzeichnet wird:

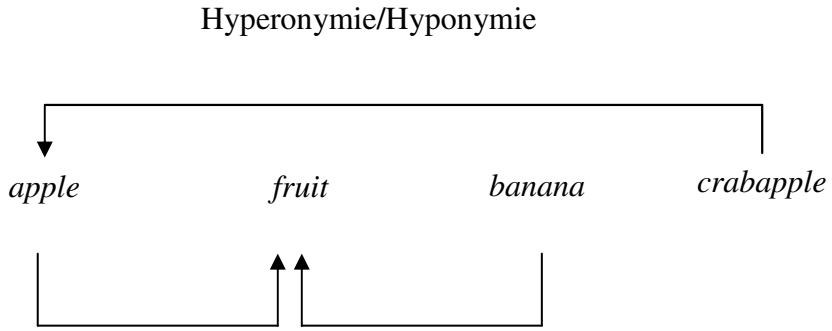


Abbildung 3-4: Darstellung einer möglichen lexikalischen Kette

Die meisten der bisherigen kohäsions-basierten Textanalysen setzen die explizite Erkennung der Kohäsionskette voraus. Dies zeigt sich auch bei der Indexierung von Stairmand und Blake, wo eine lexikalische Kette als Ganzes berücksichtigt wird und die durchschnittliche Stärke der einzelnen lexikalischen

Ketten als Gewicht aller Terme des Clusters, von dem lexikalische Ketten hergeleitet werden, bestimmt ist (vgl. Stokes 2004:34 f.). Nach dieser Strategie ist dasselbe semantische Gewicht all den Indextermen zugewiesen, die zur derselben lexikalischen Kette bzw. demselben Cluster gehören. Was in dieser Arbeit versucht wird, ist die Bestimmung der Indextermgewichtung anhand der kognitiven Spuren der lexikalischen Relationen von Wörtern. Deshalb stehen die Informationen über einzelne lexikalische Relationen im Text im Mittelpunkt. Wie Abbildung 3-4 zeigt, kann man eine erkannte lexikalische Kette als einen gerichteten Graph (vgl. Diestel 2006:45 f.) betrachten, wobei dessen Kanten anhand der lexikalischen Relationen bestimmt sind. In Bezug auf die Graphentheorie lässt sich sagen, dass ein Graph in dieser Arbeit nicht in erster Linie als Ganzes betrachtet wird, sondern die einzelnen Teilgraphen von Interesse sind, die von dem einzelnen Knoten und dessen benachbarten Knoten gebildet werden. Dies wird noch mal in Abschnitt 3.3.5 aufgegriffen.

3.3.1.1 Transitivitätsproblem

Im Abschnitt 2.1 wurde gezeigt, dass Transitivität ein wichtiges Charakteristikum der Hyponymie und der Meronymie ist, die sich zu hierarchischen Relationen anordnen lassen. Sie wirken bei der Konzipierung der kohäsions-basierten Textanalyse in zwei Hinsichten, die eng miteinander verbunden sind: Das erste Problem ist die Frage, ob Transitivität bei der Erkennung der hierarchischen Relationen in Betracht gezogen wird, und wenn ja, wie weit man sie dann berücksichtigt. Das zweite Problem bezieht sich auf die Verkettungsstrategie. In diesem Abschnitt wird das erste Problem behandelt, das zweite in Abschnitt 3.3.1.4.

Durch die Berücksichtigung der Transitivität wird es ermöglicht, die Verbindung von zwei Wörtern, die in dem lexikalischen System nicht unmittelbar miteinander zusammenhängen, zu erkennen. So kann etwa *vehicle* als Hyperonym nicht nur von *craft* erkannt werden, sondern auch von den Wörtern, die sich in dem hyponymen Baum in Abbildung 3-5 auf den unteren Ebenen befinden.

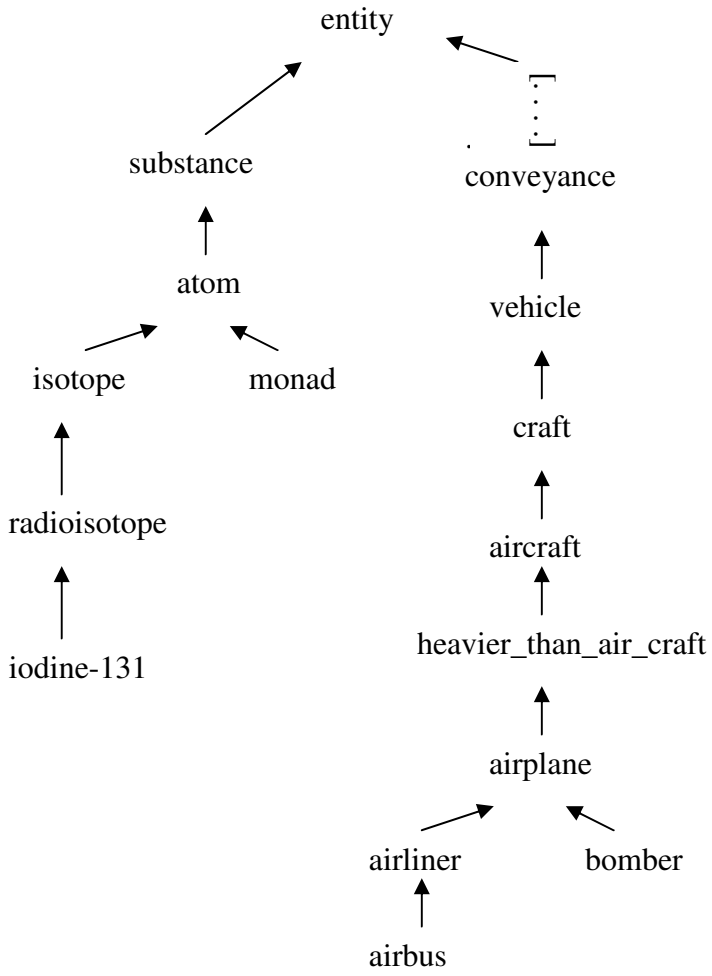


Abbildung 3-5: Zwei hyponyme Teilbäume aus WordNet

Es ist kaum vorstellbar, dass sich ein Textproduzent des gesamten lexikalischen Systems bewusst ist und die Fortführung desselben Koreferenten bzw. Konzepts durch die unmittelbar darüber stehenden Hyperonyme realisiert. Nicht selten ist es der Fall, dass die Ausdrücke des Koreferenten durch Hyperonyme realisiert sind, deren Distanz zu ihrem Hyponym sehr groß ist. Damit ist der Vorteil der Berücksichtigung der Transitivität offensichtlich. In diesem Zusammenhang führen Hirst/St. Onge (1998:308) den Begriff „Pfad“ ein, der die Distanz zwischen zwei Synsets (vgl. Abschnitt 2.2.3.2) von WordNet, zu denen beide Wörter gehören, bezeichnet und dessen Wert anhand der Anzahl der Verbindungen zwischen zwei Synsets berechnet wird. So kennzeichnet zum Beispiel die Verbindung von *substance* und *stom* die Pfadlänge 0, da in Bezug auf das Synset in WordNet, zu dem beide Wörter gehören, die unmittelbare Verbindung der Hyponymie in WordNet besteht. Und die Verbindung zwischen

bomber und *aircraft* kennzeichnet die Pfadlänge 2, da die kürzeste Verbindung von einem Wort zum anderen über fünf andere Synsets hergestellt werden kann. Das Einsetzen der Pfadlänge mit hoher Zahl führt zu der Erkennung der Verbindung zwischen den Wörtern, die sich in derselben Hierarchie befinden. Aber dieser Vorteil ist eng verbunden mit dem kritischen Aspekt, der darauf beruht, dass die Längen der hierarchischen Ketten im lexikalischen System unterschiedlich sind, sowie darauf, auf welcher Ebene in der Hierarchie sich das Wort befindet. Während alle Hyponyme in der Hierarchie, in der sich das Wort *atom* befindet, mit der Pfadlänge 4 erkannt werden können, braucht die andere Hierarchie dafür die Pfadlänge 11. Dies verursacht folgendes Problem: Wenn man von *monad* aus hierarchisch zwei Ebenen höher geht, dann erreicht man die Ebene, in der sich *substance* befindet, und man verliert die meisten semantischen Informationen von *monad*. Hingegen ist die semantische Verbundenheit zwischen *airbus* und *vehicle* immer noch intuitiv erfassbar, obwohl die Pfadlänge zwischen beiden 5 ist. Wenn die Pfadlänge mit 1 oder 2 eingestellt wird, dann kann *isotope* mit *iodine-131* verbunden werden, aber auch mit *entity*, das sehr wenig semantischen Gehalt besitzt. Daraus ist die Gefahr zu erkennen, dass zwei in hierarchischer Relation zueinander stehende Wörter nicht semantisch ähnlich sein müssen. Noch ein damit verbundener negativer Aspekt ist, dass, wenn das Wort *entity* oder *thing* vorkommt, zu viele Verbindungen zwischen diesen und anderen Wörter bestehen. Die Erweiterung der Pfadlänge bei hierarchischen lexikalischen Relationen kann die Präzision bei der Ermittlung der semantischen Verbindungen stark minimieren und wenig zum Nachweis der kognitiven Spur beitragen. Insgesamt zeigt sich die Entscheidung über das Einsetzen der Pfadlänge als sehr schwierig. Wenn man sie überhaupt einsetzen will, dann ist sie nicht allgemein bestimmbar, sondern nach den verschiedenen Faktoren wie Wortauswahlstendenz der Textproduzenten für die Ausdrücke der Koreferenz, der Repräsentationsgranularität des lexikalischen Systems, der Länge der Hierarchien, in denen sich die Wörter befinden, und der Höhe der einzelnen Wörter in den Hierarchien. Das heißt, einzelne Wörter besitzen unterschiedliche Pfadlängen je nach Text und Wissensbasis. In dieser Arbeit wird die Transitivität bei der Erkennung der hierarchischen Relationen nicht weiter berücksichtigt.

3.3.1.2 Greedy- und Non-Greedy-Algorithmen

Die meisten kritischen Probleme bei der kohäsions-basierten Textanalyse entstehen hauptsächlich daraus, dass einer Wortform mehrere mögliche Konzepte zugewiesen werden können. Deshalb kann es auch mehrere mögliche Verbindungen von einer Wortform zu anderen Wörtern durch ihre verschiedenen Konzepte geben. Die Bestimmung des passenden Konzepts zu einer Wortform und die Erstellung der lexikalischen Kette stehen unter einer Wechselbeziehung, das heißt, sie sind füreinander vorausgesetzt. Für die Verkettung bzw. Disambiguierung sind zwei Algorithmen denkbar: Während bei sog. Non-Greedy-Algorithmen alle Erstellungen oder Zuweisungen der Elemente zu den Clustern verschoben werden, bis alle Zuweisungsmöglichkeiten herausgefunden sind, basiert bei Greedy-Algorithmen die Zuweisung eines Elements zu einem Cluster auf den Elementen, die bereits diesem Cluster zugewiesen sind und nach denen die Zugehörigkeit eines neuen Elements sofort festgelegt ist (vgl. Cormen/Leiserson/Rivest/Stein 2001:329-355, Stokes 2004:28 ff.). Das heißt, wenn ein Wort im Text durch eines von vielen möglichen Konzepten, die einer Wortform zugewiesen werden können, in einem Cluster aufgenommen wird, dann ist bei Greedy-Algorithmen die Analyse des Wortes abgeschlossen, obwohl die Möglichkeit bestünde, durch weitere Analyse bessere, also längere lexikalische Ketten zu bilden. Dagegen werden bei Non-Greedy-Algorithmen erst alle möglichen konzeptuellen Verbindungen erstellt, und danach werden je nach Strategie die angemessenen Verbindungen ausgewählt. Die Auswahl zwischen den beiden Algorithmen hängt normalerweise von der erwünschten Genauigkeit und dem Aufwand der Analyse ab, wobei Vor- und Nachteile der beiden Algorithmen bekannt sind. Der Grund, warum in dieser Arbeit der Verkettungsprozess mit den Non-Greedy-Algorithmen[49] vorgenommen wird, hängt nicht unbedingt mit der Erwartung eines besseren Ergebnisses der Disambiguierung zusammen, die als Hauptvorteil der Non-Greedy-Algorithmen gilt. Der Hauptgrund liegt vielmehr in der Strategie für die Gewichtsbestimmung der Indexterme, die in Abschnitt 3.3.5 behandelt wird.

[49] Wie im Verlauf der weitern Beschreibung klar wird, ist in dieser Arbeit die explizite Bestimmung der endgültigen Konzeptcluster nicht vorgenommen, die eng mit der lexikalischen Disambiguierung verbunden ist. Dennoch ist hier wegen der Komplexität bei der Ermittlung von möglichen lexikalischen Ketten von Greedy- und Non-Greedy-Algorithmen die Rede.

3.3.1.3 Problem der Relationsdarstellung

Noch ein Problem ist zu erkennen, wenn Wiederholungen oder Synonyme im Text vorkommen, nämlich wenn mehrere Wörter, die zu demselben Synset in WordNet gehören, dieselben lexikalischen Relationen zu anderen Wörtern im Text besitzen. Dies wird durch das Beispiel in Abbildung 3-6 gezeigt, wobei *calculator* zweimal im Text vorkommt, das ein Synonym von *reckoner* ist, *statistician* und *actuary* Synonyme sind und die Synsets {*calculator*, *recockoner*} und {*statistician*, *actuary*} unter Hyponymie stehen:

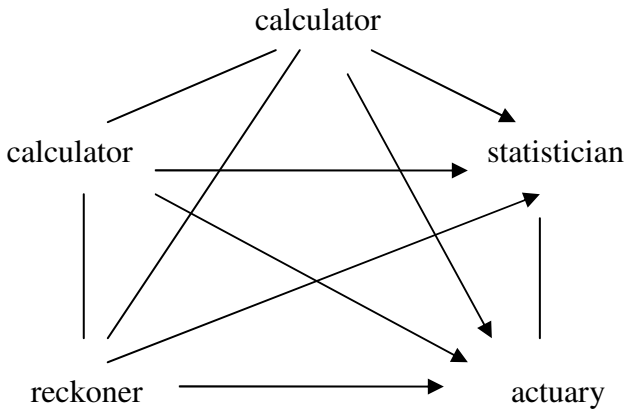


Abbildung 3-6: Eine mögliche Darstellung der lexikalischen Kohäsion

Vorher wurde erwähnt, dass Synonyme wegen ihrer gemeinsamen deskriptiven Bedeutung normalerweise mit denselben Wörtern dieselben lexikalischen Relationen eingehen. Diese Charakteristik ermöglicht die Überlegung, bei der Erkennung einer lexikalischen Kette erst die Gruppen der Synonyme und Wiederholungen und dann weitere lexikalische Relationen anhand dieser Gruppen zu ermitteln. Durch diese Überlegung können die unübersichtlichen Darstellungen der lexikalischen Ketten reduziert werden, so kann man etwa die in Abbildung 3-6 dargestellten Relationen folgendermaßen vereinfachen:

{*calculator*, *recockoner*} {*statistician*, *actuary*}

Der größte Vorteil der Gruppierung von Synonymen und Wiederholungen liegt darin, dass die Einheit, die später als Indexterm betrachtet wird, nicht nur durch formale Ähnlichkeit, sondern semantisch bzw. konzeptuell bestimmt werden

kann. Auch der Aufwand des Erkennungsprozesses kann reduziert werden, indem Synonyme und Wiederholungen als ein Konzept betrachtet werden. Man kann die Erkennung der lexikalischen Kette im Text als Ermittlung der Teilmenge der in WordNet definierten Synsets und ihrer Relationen betrachten. Diese Überlegung beeinflusst die Bestimmung der Datenstruktur und der Algorithmen für die Erkennung der lexikalischen Kette, die in den nächsten zwei Abschnitten behandelt wird.

3.3.1.4 Clustering-Algorithmen

Der Versuch, die möglichen Verbindungen von Wörtern zu anderen Wörtern zu erkennen, führt zur Überlegung, dass die Erstellung der lexikalischen Kette mit Clustering-Algorithmen (vgl. Kowalski 1997:125-148, Manning/Schütze 2000:495-528) durchgeführt werden kann. Aber die direkte Anwendung der gängigen Clustering-Algorithmen ist für die Erkennung der lexikalischen Ketten aus dem Grund nicht ganz unproblematisch, dass eine lexikalische Kette durch die verschiedenen Arten der lexikalischen Relationen dargestellt wird. Wenn man also eine lexikalische Kette mit einem Graph vergleicht, führt die Charakteristik der lexikalischen Relationen dazu, dass die Kanten im Graphen vielfältigen logischen Charakter besitzen können, was beachtet werden muss.

Zuerst betrachten wir das Single-Link-Clustering, dessen Vorgang wie folgt beschrieben ist:

> „1. Select a term that is not in a class and place it in a new class
> 2. Place in that class all other terms that are related to it
> 3. For each term entered into the class, perform step 2
> 4. When no new terms can be identified in step 2, go to step 1“(Kowalski 1997:134)

Die Zugehörigkeit eines Elements zu mehreren Klassen ist beim Single-Link-Clustering ausgeschlossen (vgl. Kowalski 1997:134). In Abschnitt 2.1 wurde gezeigt, dass sich lexikalische Relationen den logischen Relationen wie reflexive, symmetrische und transitive Relationen zuordnen lassen. Wenn eine lexikalische Kette nur durch Synonymie und Wiederholung gebildet würde, die zu den symmetrischen Relationen gehören, dann genügten die ersten zwei Schritte

des Single-Link-Algorithmus. Dies liegt daran, dass ein Wort seine Synonyme konzeptuell vertritt, und daraus ist zu erkennen, dass, wenn die Synonyme ermittelt sind, nicht jedes Synonym mit anderen Wörtern verglichen werden muss. Wenn noch eine symmetrische Relation wie Antonymie in Betracht gezogen wird, benötigt der Erkennungsprozess weiter die restlichen Schritte, also Schritt 3 und 4, wobei die konzeptuelle Eigenschaft der Synonymie immer noch nutzbar ist. Die Antonymie ist meistens konzeptuell binär, das heißt, es gibt nur zwei Konzepte, die eine Antonymie bilden. Wenn sie zwischen zwei Wörtern erkannt ist, gibt es zwei Konzepte in der aktuell zu verarbeitenden lexikalischen Kette, mit denen die anderen Wörter im Text verglichen werden müssen, und das kann der Prozess bei den weiteren Schritten des Teilprozesses für die Erkennung der Antonymie sparen. Wenn also w_1 Antonym von w_2 und Synonym/Wiederholung von w_3 ist, dann ist w_3 auch Antonym von w_2. Dies gilt nicht nur für Antonymie, sondern auch für andere lexikalische Relationen.

Kritisch wird die Berücksichtigung der hierarchischen Relationen wie Hyponymie und Meronymie, die transitiv sind. Wenn die Wörter mit der Reihenfolge *person*, *male* und *female* vorkommen, dann erkennt die Single-Link-Technik nicht die Antonymie, die zwischen *male* und *female* besteht. Dies liegt daran, dass *male* und *female* anhand des Wortes *person* als Hyponyme erkannt worden sind und es keine Möglichkeit gibt, andere Relationen zu ermitteln, die zwischen den schon zu einer Klasse gehörigen Wörtern bestehen. Um zu versuchen, dieses Problem zu lösen, wird die Clique-Technik eingeführt, die im folgenden Algorithmus dargestellt wird:

for i = 1 **to** m **do**
  lege einen neuen Cluster mit $term_i$ an
  for r: = i+1 **to** m **do**
    for k: = r **to** m **do**
      If $term_k$ zu allen Termen im aktuellen Cluster ähnlich ist
        then füge $term_k$ zum aktuellen Cluster hinzu
      lege einen neuen Cluster mit $term_i$ an
  If aktueller Cluster nur $term_i$ enthält, und $term_i$ ist auch in anderen Cluster enthalten
    then lösche den aktuellen Cluster
  lösche gleiche Cluster oder Untercluster

Clique-Algorithmus (Kowalski 1997:133 f.)

Zwei Nachteile dieses Algorithmus sind erkennbar: Erstens können zwei Elemente mehrmals verglichen werden. Dies kann dadurch gelöst werden, dass man eine Wiederholungsschleife weglässt, die für den Zweck dieser Arbeit nicht notwendig ist. Ebenfalls nicht nötig ist es, dass ein Wort mit allen anderen Wörtern, die schon in der Kette eingetragen sind, lexikalische Relationen bildet, um in die Kette aufgenommen zu werden. Aber der Analyseaufwand bleibt gleich, um mögliche lexikalische Relationen unter den Wörtern zu ermitteln. Zweitens läuft der Vergleich nur vorwärts, und das Zurücksetzen der Analyse ist nicht erlaubt, das für die Erkennung der hierarchischen Ketten abhängig von der Vorkommensreihenfolge nötig ist. Wenn beispielsweise die Wörter in der Reihenfolge *crabapple*, *fruit* und *apple* vorkommen (und keine Transitivität berücksichtigt wird), dann wird *fruit* nicht ohne Rücklauf des Analysevorgangs in der Kette aufgenommen. Dasselbe Problem tritt auch auf, wenn die Wörter in der Reihenfolge *isotope … substance … atom* vorkommen, wobei – wie Abbildung 3-5 zeigt – *isotope* und *substance* in WordNet nicht als unmittelbare Hyponyme definiert sind. Ein ähnliches Problem liegt vor, wenn die Wörter in der Reihenfolge *apple*, *banana* und *fruit* stehen und Kohyponymie beim sequenziellen Vergleich nicht in Betracht gezogen wird. Dann werden nur die Relationen zwischen *apple* und *fruit* in derselben Kette bzw. derselben Klasse erkannt. In diesen Fällen gehören dieselben Wörter zu verschiedenen Klassen bzw. Ketten. Sicherlich ist es keine absolute Voraussetzung für die Berechnung der semantischen Gewichte einzelner Konzepte, alle Wörter, die zu demselben hierarchischen Baum gehören, in derselben Klasse zu erkennen. Man könnte nach dem Abschluss der Erkennung aller möglichen Klassen das semantische Gewicht der Wörter berechnen. Aber da die semantische Gewichtung von Wörtern nach ihren lexikalischen Verbindungen zu anderen Wörtern berechnet wird, ist der Vorteil ohne Schwierigkeit erkennbar, wenn man die Wörter, unter denen lexikalische Relationen bestehen, derselben Klasse zuweist. Aus den bisherigen Überlegungen wird deutlich, dass es unangemessen ist, beide Clustering-Algorithmen für den Zweck dieser Arbeit ohne zusätzliche Vorkehrungen anzuwenden, die die erwähnten Probleme abdecken. In Abschnitt 3.3.3 wird versucht, einen Prozess für die Erkennung der lexikalischen Kette zu bestimmen, dessen Grundidee auf die Single-Link-Technik zurückzuführen ist, wobei folgende Faktoren berücksichtigt werden: Erstens muss es die Möglich-

keit geben, ein Element mit den anderen Elementen in derselben Klasse zu vergleichen. Zweitens sollten sich Synonyme bei der Kettenbildung in einer Gruppe in derselben Datenstruktur eintragen lassen. Drittens sollte jedes Wort mit anderen Wörtern nur einmal verglichen werden.

3.3.2 Datenstruktur

Für die Indexierung werden zwei grundlegende Datenstrukturen eingeführt: Die erste wird „LUinfo“ genannt und entspricht einem einzelnen Vorkommen der lexikalischen Einheit im Text, die die Kombination von Wortform und einem möglichen der Konzepte enthält, die einer Wortform zugewiesen werden können. Jedes Token im Text wird als ein Lemma betrachtet, und dadurch entsteht eine Anzahl von n LUinfo-Strukturen bei jedem Token, wobei n die Anzahl der möglichen Belegformen bzw. Konzepte ist. Ein Konzept wird mit der Dateiadresse, bei der der Eintrag des entsprechenden Synsets in der Datendatei in WordNet anfängt, identifiziert. Neben Konzept und Wortform enthält diese Datenstruktur Informationen über die Wortklasse und die Stelle, an der das Token im Text auftritt. Demzufolge sieht die LUinfo-Struktur folgendermaßen aus:

```
LUinfo{
      char *wordform;
      int pos;
      long addr;
      int word_number;
     int  processed;
 };
```

Während die LUinfo-Struktur möglichen Interpretationen zu dem im Text vorkommenden einzelnen Wort entspricht, stellt die ICinfo-Struktur dasselbe Konzept dar, das im Text an unterschiedlichen Stellen zum Ausdruck gebracht wird. LUinfo- und ICinfo-Struktur sind voneinander abhängig, das heißt, eine LUinfo-Struktur kann nur bestehen, indem sie in einer ICinfo-Struktur eingetragen ist, und eine ICinfo-Struktur kann ohne einen einzigen Eintrag der LUinfo-Struktur nicht existieren:

```
ICinfo{
     int icId;
```

```
    LUinfo **luContainer;
    long weight;

    int  **hyperonymie;
    int  **hyponymie;
    int  **antonymie;
    int  **ursache-wirkung;
    int  **meronymie_index;
    int  **holonymie_index;
    int  ** derivation;
};
```

Wenn eine LUinfo-Struktur erstellt wird, wird sie entweder in der vorhandenen ICinfo-Struktur eingetragen, oder eine neue ICinfo-Struktur wird erzeugt, die sie aufnimmt. Dies hängt mit der ursprünglichen Aufgabe der ICinfo-Struktur zusammen, lexikalische Relationen darzustellen. Wenn die durch die LUinfo dargestellten Informationen zum gleichen Synset in WordNet gehören, mit anderen Worten, wenn sie unter Wiederholung oder Synonymie stehen, werden sie in derselben ICinfo-Struktur eingetragen. Synonymie und Wiederholungen werden ohne explizite Verbindungsinformationen nur durch den Eintrag in derselben ICinfo dargestellt. Weitere lexikalische Verbindungen beruhen auf den ICinfo-Strukturen, indem sie untereinander zur Verbindung gebracht werden. Deswegen benötigt normalerweise eine lexikalische Kette mehrere ICinfo-Strukturen, kann aber auch lediglich eine ICinfo-Struktur besitzen, wenn die Kette nur aus Wiederholung und Synonymie besteht.

3.3.3 Erkennungsalgorithmus

Für die Vereinfachung der Darstellung des Algorithmus wird hier angenommen, dass jeder Wortform nur ein passendes Konzept zugewiesen ist. Das heißt, die Variable lu bezeichnet eine lexikalische Einheit, die einer LUinfo-Struktur entspricht. In der Implementierung wird die Möglichkeit berücksichtigt, dass mehrere LUinfo-Strukturen aus einem Typ entstehen. Die Wörter, die in diesem Abschnitt als ein Beispiel für einen Verkettungsprozess verwendet werden, werden durch Tabelle 3-1 dargestellt, und Tabelle 3-2 zeigt die in WordNet definierten lexikalischen Relationen.

lu_1	lu_2	lu_3	lu_4	lu_5	lu_6	lu_7	lu_8
girl	person	female	girl	male	male	man	female

Tabelle 3-1: Wörter im Text

Wiederholung/ Synonymie(0)	Antonymie(1)	Hyperonymie(2)	Hyponymie(3)
girl – girl	female – male	female – girl	girl – female
female – female		person – female	female – person
male – male		person –male	male – person
		male – man	man – male

Tabelle 3-2: Darstellung der lexikalischen Relationen

Die Nummern 0-3 hinter den Bezeichnungen der lexikalischen Relationen werden später in den Tabellen wieder verwendet, die die Prozessschritte des Algorithmus darstellen. Wenn man versucht, mit den in der Tabelle 3-1 genannten LUinfo-Strukturen nach dem Single-Link-Algorithmus, der dem Verkettungsverfahren dieser Arbeit zugrunde liegt, eine Kette zu bilden, dann wird zuerst lu_1 als Bezugselement bestimmt anschließend wird durch einen Vergleich geprüft, ob lu_2 bis lu_8 aufgrund lexikalischer Relationen mit lu_1 in dieselbe Kette aufgenommen werden können. Nach diesem Prozess entsteht eine Kette, die aus lu_1, lu_3, lu_4, lu_8 besteht, die jeweils *girl*, *female*, *girl*, *female* entsprechen. Anschließend wird die unmittelbar nach dem aktuellen Bezugselement lu_1 in die Kette aufgenommene LUinfo-Struktur, also in diesem Fall lu_3, als ein neues Bezugselement bestimmt, und dieses wird weiter mit lu_2, lu_5, lu_6 und lu_7 verglichen, die wiederum in die Kette aufgenommen werden, wenn sie mit lu_3 lexikalische Relationen bilden. Dies wiederholt sich, bis kein Element mehr vorhanden ist, das als Bezugselement dient. Für die weitere Erklärung wird eine Variable lu_j definiert, die durch die Bezugselemente instanziiert werden kann, und die möglichen LUinfo-Strukturen, die mit lu_j verglichen werden, werden als lu_k dargestellt. Demzufolge bezieht sich lu_j beim Prozess der Beispielelemente nach dem Single-Link-Algorithmus zuerst auf lu_1, lu_k zuerst auf lu_2. Weiter werden lu_3 bis lu_8 nach der Reihenfolge als lu_k instanziiert. Danach wird die unmittelbar nach der aktuellen lu_j aufgenommene lu_k als neue lu_j, also lu_3 bestimmt, deren lu_k die Elemente benötigt, die noch nicht in die Kette aufgenommen worden sind. Die Art der Bestimmung von lu_j und lu_k wird bei dem

in dieser Arbeit verwendeten Verkettungsverfahren wegen der Charakteristik der Relationen, nach der lu_l und lu_k verglichen werden, modifiziert.

Vor der Beschreibung der Verkettungsverfahren ist es nötig, die zwei verwendeten Vergleichsoperationen zu erläutern: „S- und O-Vergleich“. Ersteres bezeichnet nur den Vergleich zwischen zwei LUinfo-Strukturen, der ermittelt, ob diese unter Synonymie oder Wiederholung stehen. Durch den O-Vergleich wird ermittelt, ob zwei Wörter unter einer der anderen hier untersuchten lexikalischen Relationen stehen. Also bezeichnen „S-Relationen“ Wiederholung und Synonymie und „O-Relationen“ alle anderen Relationen außer S-Relationen, nämlich Antonymie, Hyperonymie, Hyponymie, Holonymie, Meronymie und Derivation. Die Motivation dieser Unterscheidung liegt darin, dass LUinfo-Strukturen, die unter einer S-Relation stehen, einer ICinfo-Struktur zugewiesen werden. Wenn also lu_l als lu_j bestimmt und mit einer lu_k verglichen wird, dann wird zuerst ermittelt, ob eine S-Relation zwischen den beiden besteht. Das ist eine Art von S-Vergleich. Wenn dies der Fall ist, dann wird die lu_k in die ICinfo-Struktur aufgenommen, in der lu_j eingetragen ist. Wenn dies nicht der Fall ist, dann erfolgt der O-Vergleich. Da davon ausgegangen wird, dass nur eine lexikalische Relation zwischen zwei Wörtern besteht, wird der Vergleichsvorgang zwischen lu_l und einer lu_k abgebrochen, sobald eine lexikalische Relation (entweder S- oder O-Relation) gefunden wurde. Zu beachten ist: Nur wenn lu_l als lu_j dient, kann der O-Vergleich unmittelbar nach einem negativen Ergebnis des S-Vergleichs zum Einsatz kommen. Wenn lu_j nicht mit lu_l instanziiert ist, dann erfolgt der O-Vergleich von lu_j nicht direkt nach einem S-Vergleich, sondern die beiden Vergleiche werden auf unterschiedlichen Prozessstufen durchgeführt. Wie später gezeigt wird, kommt in der in dieser Arbeit vorgestellten Verfahrensweise auch der Zustand vor, in dem eine lu_j nur ihren S-Vergleich abgeschlossen hat. Dieser Zustand wird als „S-Vergleichabschluss“ bezeichnet, und lu_j wird mit ihren lu_k nur noch auf O-Relationen geprüft. Von „Vergleichsabschluss“ oder „Vergleich abschließen“ einer lu_j ist immer dann die Rede, wenn sie erst mit allen möglichen LUinfo-Stukturen, die als lu_k dienen, auf S- und O-Relation geprüft wurde.

Das Verkettungsverfahren dieser Arbeit unterscheidet sich von dem Single-Link-Algorithmus vor allem dadurch, dass es die Möglichkeit berücksichtigt wird, unter den lu_k zu vergleichen, die durch ihre Relationen zu einer lu_j in die Kette aufgenommen wurden. Also wird lu_k unabhängig von der Kettenzugehörigkeit der LUinfo-Strukturen bestimmt. Eine LUinfo-Struktur wird als lu_k bestimmt, wenn ihr Vergleich nicht komplett abgeschlossen ist. Dieser Fall tritt auch ein, wenn nur der S-Vergleich in Bezug auf lu_k abgeschlossen ist. Wie eine ICinfo in den Zustand des S-Vergleichabschlusses gelangt, hängt mit dem zweiten Faktor zusammen, der sich auf dem in Abschnitt 3.3.1.3 erwähnten Darstellungsaspekt bezieht. Es werden nämlich die LUinfo, die unter einer S-Relation stehen, derselben Gruppe, also derselben ICinfo-struktur, zugewiesen. Dies beeinflusst die Bestimmung der nächsten lu_j und lu_k. Wenn sich eine LUinfo in der schon vorhandenen ICinfo hinzufügen lässt, kann man sie bei der Bestimmung der nächsten lu_j weglassen. Wenn zum Beispiel lu_1 (*girl*) und lu_4 (*girl*) jeweils als lu_j und lu_k bei einem Vergleich bestimmt sind, dann wird lu_4 in die ICinfo-Struktur eingetragen, in der lu_1 eingetragen ist; man braucht später nicht lu_4 als lu_j zu bestimmen, da lu_1 lu_4 konzeptuell vertritt. Ähnlich gilt dies auch bei der Bestimmung einer lu_k. Wenn lu_1 und lu_4 jeweils als lu_k zu der gemeinsamen lu_j bestimmt sind und durch die Wiederholung in einer ICinfo eingetragen sind, dann kann nur eine von den beiden LUinfo bei der Bestimmung der nächsten lu_k, die sich auf die andere lu_j bezieht, in Betracht gezogen werden. Die Instanziierung der lu_j geschieht nicht nach der Reihenfolge der in der Kette aufgenommenen LUinfo-Strukturen, sondern nach der Reihenfolge der Aufnahme der ICinfo-Strukturen in die Kette. Dies zeigt einen positiven Aspekt der Gruppierung der Wörter, die unter S-Relation stehen. Je mehr also die unter S-Relation stehenden Wörter im Text vorkommen, desto geringer wird der Verkettungsaufwand. Problematisch ist, dass eine S-Relation nicht nur zwischen einer lu_j und einer lu_k, sondern auch zwischen der aktuellen lu_k und früheren lu_k bestehen kann, die durch eine O-Relation in die Kette aufgenommen wurden. Also besteht eine S-Relation auch zwischen lu_3 (*female*) und lu_8 (*female*), die zudem beide durch die O-Relation Hyponymie mit lu_1 (*girl*) verbunden sind. Die Einführung einer weiteren Variable kann hilfreich für die weitere Darstellung sein: lu_r bezeichnet eine Variable für die vorherigen lu_k, die sich gemeinsam mit der aktuellen lu_k auf die gleiche lu_j beziehen und durch die

O-Relation in die Kette aufgenommen wurden. Die S-Relation zwischen der aktuellen lu_k und der lu_r könnten dadurch erkannt werden, dass die früher in der Kette aufgenommenen lu_k – also die lu_r – später als lu_j dienen. In diesem Fall ist der Nachteil zu erkennen, dass eine zusätzliche Vorrichtung für die Verwaltung der ICinfo-Strukturen benötigt wird, in der die LUinfo-Strukturen eingetragen sind, die unter S-Relation stehen. Also muss man zwei ICinfo-Strukturen, die dasselbe Konzept darstellen, in eine ICinfo zusammenführen. Um dies zu vermeiden, nimmt das in dieser Arbeit verwendete Verkettungsverfahren die folgende Vorkehrung vor: Wenn eine aktuelle lu_k mit der lu_j eine O-Relation bildet, wird vor ihrem Eintrag in der Kette ermittelt, ob sich eine lu_r durch ihre entsprechende ICinfo in der Kette befindet, die dasselbe Konzept wie die aktuelle lu_k darstellt. Es wird also ein S-Vergleich zwischen der aktuellen lu_k und den lu_r durchgeführt. Wenn solch ein S-Vergleich positiv ist, dann wird die aktuelle lu_k in derselben ICinfo eingetragen, und man braucht sich nicht extra um die Darstellung der O-Relation zwischen der aktuellen lu_k und der lu_j zu kümmern, da diese O-Relation schon vorher in der ICinfo markiert ist, in der die entsprechende lu_r eingetragen wurde. Es ist zu erkennen, dass, wenn der Vergleich einer lu_j abgeschlossen ist, sie also mit allen anderen lu_k verglichen wurde, auch die S-Vergleiche aller LUinfo abgeschlossen sind, die als lu_k durch O-Relationen zu der gemeinsamen lu_j in der Kette aufgenommen wurden. Dies führt dazu, dass S-Vergleiche gespart werden können, wenn lu_k später als lu_j dient. Im Prinzip bezieht sich ein S-Vergleich auf die aktuellen lu_k und lu_r, da die meisten lu_j vorher als lu_k gedient haben. Es gibt nur eine Instanz der lu_j, die vorher nicht als lu_k gedient hat, nämlich lu_1. Wenn eine LUinfo, die als die allererste lu_j instanziiert ist, also lu_1 als lu_j betrachtet wird, dann ist lu_1 in dem Zustand, in dem weder der S- noch der O-Vergleich abgeschlossen ist. Außer lu_1 sind alle lu_j automatisch beim S-Vergleich abgeschlossen. Das heißt, wenn lu_1 als lu_j instanziiert wird, dann werden beim Vergleich zwischen lu_j und lu_k im Prinzip die S- und die O-Relationen geprüft, ansonsten werden nur die O-Relationen geprüft.

Die Tabellen 3-3 bis 3-6 und die Abbildungen 3-7 bis 3-11 stellen jeweils den Ablauf des Verkettungsprozess während der Erkennung der lexikalischen Kette dar, die aus den LUinfo besteht, die in Tabelle 3-1 dargestellt werden. Tabelle

3-3 veranschaulicht den Verkettungsprozess, wenn lu_1 als lu_j bestimmt ist. Zuerst wird eine neue ICinfo erzeugt für lu_1, die als lu_j bestimmt ist, und mit lu_2 bis lu_8 verglichen, die lu_k entsprechen.

P-Nr.	lu_j – lu_k	Vorgenommene Vergleichsart	lu_k – lu_r	R-Nr.
0-0	lu_1	S,O		-
1-0	lu_1 (girl) – lu_2 (person)	S,O		
2-0	lu_1 (girl) – lu_3 (female)	S,O		3
3-0	lu_1 (girl) – lu_4 (girl)	S,O		0
4-0	lu_1 (girl) – lu_5 (male)	S,O		
5-0	lu_1 (girl) – lu_6 (male)	S,O		
6-0	lu_1 (girl) – lu_7 (man)	S,O		
7-0	lu_1 (girl) – lu_8 (female)	S,O		3
7-1		S	lu_8 (female) – lu_3 (female)	0

Tabelle- 3-3: Vergleich der lu_1

P-Nr. (Prozessnummer) 2-0 zeigt, dass *girl* das Hyponym von *female* ist. und es wird eine neue ICinfo ic_2 erzeugt, in der lu_3 eingetragen wird. Abbildung 3-7 stellt den Kettenzustand nach P-Nr. 2-0 dar, wobei eine Ellipse eine ICinfo bezeichnet. P-Nr. 3-0 zeigt den Fall, dass eine S-Relation erkannt wird, und lu_4 wird in ic_1 eingetragen, in der auch lu_1 eingetragen ist. Der Zustand nach P-Nr. 3-0 wird in Abbildung 3-8 gezeigt, in der zwei ICinfo ic_1 und ic_2 und ihre Verbindung zu sehen sind.

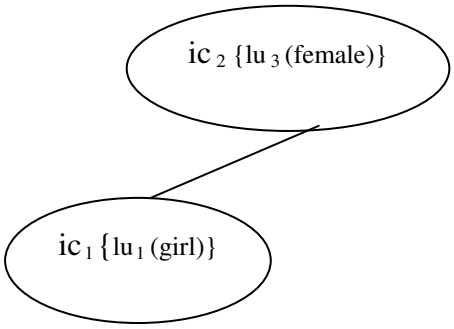


Abbildung 3-7: Nach P-Nr. 2-0

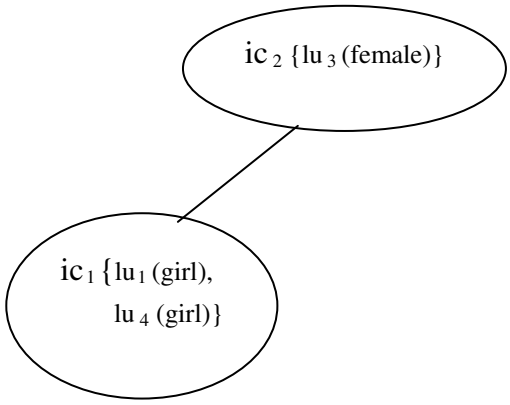


Abbildung 3-8: Nach P-Nr. 3-0

P-Nr. 7-1 zeigt den Fall, nach dem eine O-Relation zwischen lu_j und lu_k erkannt wurde. Die lu_k ist noch nicht endgültig in der Kette eingetragen, und es soll ermittelt werden, ob sich schon eine ICinfo in der Kette befindet, die dasselbe Konzept wie lu_k darstellt. Da die ICinfo, die lu_3 enthält, dasselbe Konzept wie lu_8 darstellt, wird lu_8 in der ICinfo eingetragen, in der lu_3 eingetragen ist. Der Zustand der Kette nach dem Prozess P-Nr. 7-1 stellt sich so wie in Abbildung 3-9 dar, wobei die grau markierte Ellipse den Vergleichsabschluss von den LUinfo angibt, die sich in der entsprechenden ICinfo befinden:

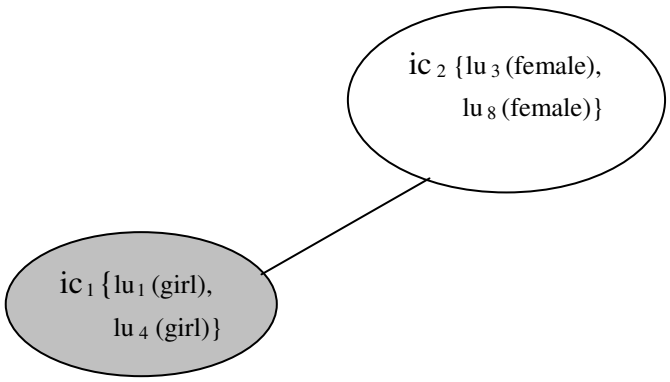


Abbildung 3-9: Nach P-Nr. 7-1

Da ein Vergleich einer lu_j dadurch geschieht, dass sie in eine entsprechende ICinfo-Struktur eingetragen wird, bedeutet ein Analyseabschluss einer LUinfo-Struktur immer den Abschluss der Analyse der entsprechenden ICinfo-Struktur, die mehrere LUinfo-Strukturen enthalten kann, die dasselbe Konzept darstellen. Daher betrachtet man den Zustand der lu_1 und lu_4 auch als Vergleichsabschluss. Das heißt, lu_4 wird bei der Bestimmung der nächsten lu_j und lu_k nicht berücksichtigt. Die nächste lu_j wird aus der ersten LUinfo der ICinfo bestimmt, die unmittelbar nach der aktuellen lu_j zugehörigen ICinfo in der Kette eingetragen ist, also ist die lu_3 aus ic_2 die nächste lu_j. Die Tabelle 3-4 stellt den Vergleich der lu_3 mit den übrigen lu_k dar:

P-Nr.	lu_j – lu_k	Vorgenommene Vergleichsart	lu_k – lu_r	R-Nr.
8-0	lu 3 (female) – lu 2 (person)	O		3
9-0	lu 3 (female) – lu 5 (male)	O		1

9-1		S	lu_5 (male) – lu_2 (person)	
10-0	lu_3 (female) – lu_6 (male)	O		1
10-1		S	Lu_6 (male) – lu_2 (person)	
10-2		S	Lu_6 (male) – lu_5 (male)	0
11-0	lu_3 (female) – lu_7 (man)	S		

Tabelle 3-4: Vergleich der lu_3

P-Nr. 8-0 und 9-0 zeigen, dass die Hyponymie zwischen *person* und *female* erkannt ist, und es wird die Antonymie gefunden, die zwischen *female* und *male* besteht. Da keine ICinfo vorhanden ist, die dieselben Konzepte wie die beiden LUinfo beinhaltet, werden die neuen ICinfo ic_3 und ic_4 erzeugt, in der lu_2 und lu_5 eingetragen werden. P-Nr. 10-0 bis 10-2 zeigen den Prozess, bei dem die lu_6 (male) in der ic_3 eingetragen wird. Dieser Prozess ist derselbe wie P-Nr. 7-1. Abbildung 3-10 stellt den Zustand der Kette nach P-Nr. 11-0 dar.

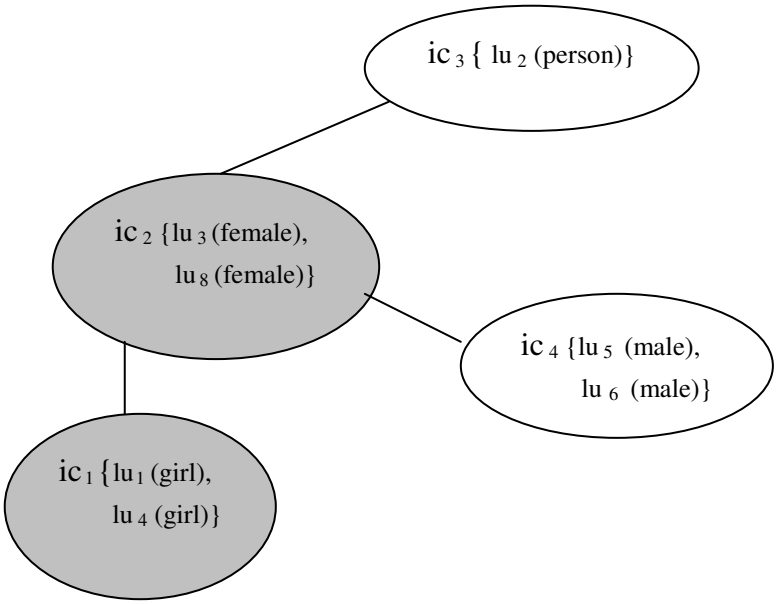


Abbildung 3-10: Nach P-Nr. 11-0

Der nächste Prozess, der in Tabelle 3-5 gezeigt wird, fängt mit der Bestimmung der lu_2 als luj an, da sie die erste LUinfo der ICinfo ic_3 ist, die unmittelbar nach der letzten lu_j zugehörigen ICinfo in der Kette eingetragen ist. Von den lu_5, lu_6 und lu_7, deren Vergleiche nicht abgeschlossen sind, werden lu_5 und

lu_7 als lu_k bestimmt, da lu_5 und lu_6 zu derselben ICinfo gehören. Das Bestimmungsprinzip der lu_j ist gleich geltend bei der Bestimmung der lu_k.

P-Nr	lu_j – lu_k	Vorgenommene Vergleichsart	lu_k – lu_r	R-Nr.
12-0	lu_2 (person) – lu_5 (male)	O		2
13-0	lu_2 (person) – lu_7 (man)	O		

Tabelle 3-5: Vergleich der lu_2

Nach dem Prozess P-Nr. 12-0, in dem ein Hyperonym erkannt ist, erfolgt kein S-Vergleich, da lu_5 in dem Zustand des „S-Vergleichabschlusses“ ist, weil lu_5 schon als lu_k gedient hat, als lu_3 die Rolle der lu_j spielte. In diesem Moment gibt keine vorherige lu_k, die sich auf die aktuelle lu_j bezieht. Deswegen gibt es in P-Nr. 12 keine lu_r. Der Verkettungszustand nach dem Schritt P-Nr. 13-0 unterscheidet sich von Abbildung 3-10 dadurch, dass die Verbindung zwischen der ic_3 und der ic_4 hergestellt und der Vergleich der lu_2 abgeschlossen ist.

P-Nr	lu_j – lu_k	Vorgenommene Vergleichsart	lu_k – lu_r	R-Nr.
14-0	lu_5 (male) – lu_7 (man)	O		2
15-0	lu_7 (man)	-		

Tabelle 3-6: Vergleich der lu_5

In der Kette bleibt eine ICinfo, deren LUinfo ihren Vergleich nicht abgeschlossen hat. Die lu_5 der ic_4 wird zu lu_j und mit lu_7 verglichen, die die einzige LUinfo ist, die nicht in die Kette aufgenommen ist. In diesem Fall existiert auch keine LUinfo, die die Rolle der lu_r übernehmen kann. Deswegen wird eine ICinfo ic_5 für die lu_5 erzeugt, und eine Verbindung zwischen ic_4 und ic_5 wird hergestellt. In Prozess P-Nr. 14-0 kommt noch eine LUinfo lu_7 hinzu, die durch ic_5 in die Kette eingetragen wird. lu_7 wird zur nächsten lu_j, aber es gibt keine LUinfo, die als lu_k dienen kann. Dies führt den Vergleichsabschluss der lu_7 herbei, und zugleich endet der ganze Prozess für ein Verkettungsverfahren. Schließlich hat

sich eine vollständige lexikalische Kette gebildet, die in Abbildung 3-11 dargestellt wird.

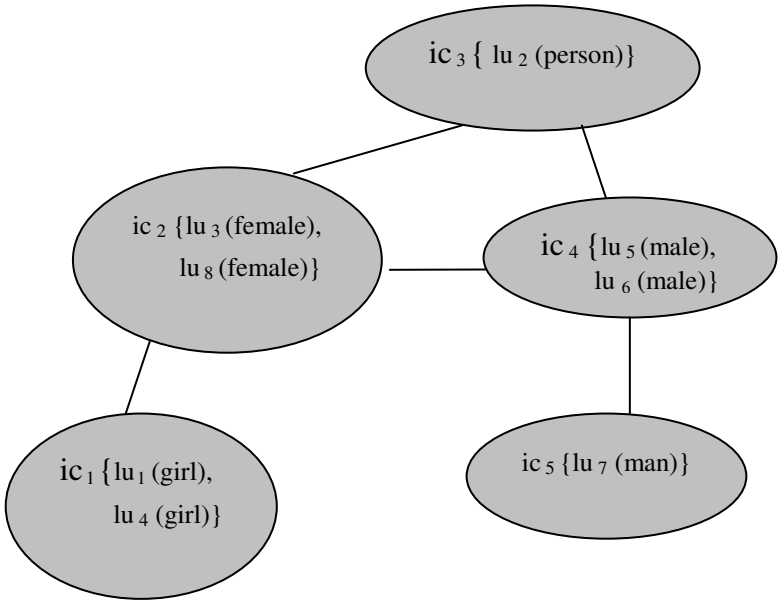


Abbildung 3-11: Nach P-Nr. 15-0

3.3.4 Bestimmung des semantischen Gewichts von Konzepten

Im theoretischen Teil dieser Arbeit wurde untersucht, welche kognitiven Funktionen durch die lexikalischen Relationen reflektiert werden können. Anhand dessen wird in diesem Abschnitt versucht, semantische Gewichte von Konzepten zu bestimmen. Nach der Erkennung der lexikalischen Kette lassen sich zwei verschiedene Arten unterscheiden, wie man die Anzahl der lexikalischen Relationen berücksichtigen kann, die sich auf die unterschiedlichen Datenstrukturen beziehen: Erstens kann man die Anzahl einer Relation aus der ganzen Anzahl der LUinfo-Strukturen berechnen, die sich in den ICinfo-Strukturen befinden, die mit der in Frage stehenden ICinfo-Struktur die lexikalische Relation bilden. Dies entspricht der Gewichtungsstrategie von Hovy/Lin (1999): Nach dieser wird das semantische Gewicht eines Konzepts durch die Summe seiner Häufigkeit und der Häufigkeit aller Vorkommensfälle seiner Hyponyme berechnet. Diese Betrachtungsweise der Anzahl einer lexikalischen Relation ist token-bezogen. Zweitens kann mit einer Anzahl einer lexikalischen Relation die Anzahl der ICinfo-Strukturen gemeint sein, die mit der gerade betrachteten

ICinfo-Struktur in bestimmten Relationen stehen. Zum Beispiel beträgt im Fall von *person* im Beispieltext aus dem vorherigen Abschnitt die Anzahl der Hyponyme 2, weil zwei ICinfo-Strukturen, die jeweils *male* und *female* entsprechen, mit der ICinfo-Struktur verbunden sind, die *person* entspricht. Wenn *male* und *female* mehrmals im Text vorkämen, bliebe die Anzahl der Hyponyme von *person* unveränderlich, da die Anzahl der ICinfo-Strukturen für ihre Darstellung nicht veränderlich ist. Also wird die Vorkommenshäufigkeit der Synonyme und Wiederholungen des Wortes w_1, mit dem das Wort w_2 in einer anderen Relation als Synonymie und Wiederholung steht, bei der Berechnung der Anzahl lexikalischer Relationen von w_2 nicht berücksichtigt. Diese Betrachtungsweise ist type-bezogen. In dieser Arbeit werden beide genannte Betrachtungsweisen zur Anzahl der lexikalischen Relationen je nach Relationsart bei der Bestimmung des semantischen Gewichts berücksichtigt.

Synonymie und Wiederholung besitzen ein hohes Potenzial an Koreferenz. Dies hängt eng mit der Steuerung der Aufmerksamkeit bei der kognitiven Textverarbeitung zusammen. Wenn der Textproduzent sich dauerhaft auf bestimmte Objekte bzw. Typen von Objekten konzentriert, bedeutet dies vermutlich, dass er davon ausgeht, dass die Objekte auch für den Rezipienten in diesem Kontext von Bedeutung sind. Je öfter ein Wort im Text erwähnt wird, desto eher ist eine semantische Relevanz durch seinen Beitrag im Text zu erwarten. Diesen kognitiven Aspekt kann man auch auf die konventionellen statistischen Termgewichtungen beziehen, wobei hier Wörtern, die im Text wiederholt auftreten, ein höherer Wert zugewiesen wird. Also deutet die Anzahl der Synonymien und Wiederholungen auf die Anzahl der LUinfo-Strukturen, die in der entsprechenden ICinfo-Struktur gespeichert sind. Konkret wird in dieser Arbeit beim ersten Vorkommen der Belegform der Grundwert 1 zugewiesen, ab dem zweiten Vorkommen durch Wiederholung und Synonymie wird dieser Wert jeweils um 0,5 erhöht.

Während Synonyme und Wiederholungen durch die Fortführung desselben Konzepts die semantische Relevanz eines Konzepts im Text nahelegen, ist dies bei Antonymie nicht im gleichen Maße der Fall, da sie ihren semantischen Charakter mehr durch den Kontrast zu anderen Konzepten deutlich werden

lässt. Obwohl der Kontext klarer wird, wenn mehrere Antonyme im Text vorkommen, sind ihre unmittelbaren Beiträge zur semantischen Relevanz des betreffenden Konzepts geringer als die von Synonymen und Wiederholungen. Dies motiviert dazu, die Anzahl der Antonyme eines Wortes bei der Synonymie und Wiederholung geringer zu gewichten. Bei der Gewichtung der Antonyme wird die Anzahl der Antonyme type-bezogen berücksichtigt. Das heißt, nur die Anzahl der ICinfo-Strukturen wird bei der Gewichtung einbezogen. Also ist das semantische Gewicht der Antonyme mit der Anzahl der ICinfo-Strukturen gleichgesetzt, unabhängig von der Anzahl der in diesen ICinfo-Strukturen enthaltenen LUinfo-Strukturen, die Antonyme darstellen. Da Antonymie meistens binär ist, ist der Fall sehr selten zu erwarten, dass mehr als zwei ICinfo-Strukturen vorkommen, die zu einem Wort Antonyme darstellen. Dieser Aspekt ermöglicht weiter die Überlegung, ob Antonymie bei der kohäsionsbasierten Textanalyse wegen ihres relativ geringen Beitrags zur semantischen Gewichtung ausgelassen werden kann.

In Abschnitt 2.3.4.2 wurde angenommen, dass Makro- und Mikrostruktur bzw. hierarchische und sequenzielle Struktur von Texten zu unterscheiden sind und einige lexikalische Kohäsionen wie Hyponymie, Meronymie und Ursache-Wirkung, die eigentlich auf dem sequenziellen Aspekt des Textes beruhen, Makro- bzw. hierarchische Textstruktur reflektieren können. Wenn eine hierarchische Struktur durch lexikalische Relationen wie Hyponymie, Meronymie oder kausale Relationen im Text realisiert wird, kann man von folgender Annahme ausgehen: Je höher sich ein Wort befindet, desto größer ist sein semantisches Gewicht in Bezug auf den reduktiven Charakter, und das auf der höchsten Ebene stehende Konzept zeigt am meisten reduktiven Charakter, da es theoretisch die auf den tieferen Ebenen stehenden Konzepte zusammenfassend repräsentiert. Weiter könnte der reduktive Charakter bei der Bestimmung des semantischen Gewichts eines Konzepts K_1 zum Beispiel dadurch berücksichtigt werden, dass man die Anzahl der Konzepte zum Konzept K_1 zuweist, die sich in der Hierarchie unterhalb von K_1 befinden. Dementsprechend haben K_2 und K_3 aus dem Beispiel in Abbildung 3-12 als ihr Gewicht jeweils 1 und 3 zugewiesen bekommen, K_1 hingegen ein Gewicht von 5.

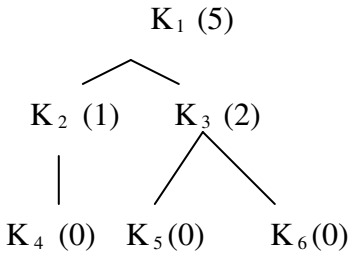


Abbildung 3-12: Eine Möglichkeit der Gewichtung eines Konzepts

Kritisch bei dieser Zuweisungsmethode ist, dass ein Konzept mit wenig semantischem Gehalt den höchsten Wert besitzen kann. Ein ähnliches Problem wurde schon erkannt, als diskutiert wurde, die Transitivität bei der Erkennung der lexikalischen Kette zu berücksichtigen (vgl. Abschnitt 2.1). Um dieses Problem zu verringern, wird die Anzahl von Konzepten bei der Gewichtung in Betracht gezogen, die sich genau eine Ebene unter dem betreffenden Konzept befinden. Demzufolge bleibt im obigen Fall das Gewicht von K_2 und K_3 unveränderlich, das von K_1 dagegen reduziert sich auf 2.

Ein Wort kann in Bezug auf seine untergeordneten Wörter eine Funktion der Zusammenfassung oder Reduktion darstellen, und je mehr die Anzahl der Meronyme, Hyponyme und die Ursache darstellenden Konzepte zunehmen, desto stärker ist die Vertretungsrolle des Konzepts. Der reduktive Charakter eines Wortes wird stärker, wenn verschiedene untergeordnete Konzepte mehrmals vorkommen, als wenn nur ein untergeordnetes Konzept mehrmals vorkommt. Aus diesem Grund wird die Häufigkeit einer hierarchischen Relation als die Anzahl der ICinfo-Strukturen bestimmt, die die untergeordneten Konzepte darstellen. Die Anzahl der LUinfo-Strukturen, die in den einzelnen ICinfo-Strukturen eingetragen sind, wird also bei der semantischen Gewichtung der hierarchischen Relationen außer Acht gelassen. Die reduktiven Eigenschaften der hierarchischen Relationen werden dadurch berücksichtigt, dass, wenn eine neue hierarchische Relation mit einer neuen ICinfo-Struktur entsteht, der Wert 1 zu den beiden ICinfo-Strukturen, die ein übergeordnetes und ein untergeordnetes Konzept darstellen, ungeachtet der Anzahl ihrer LUinfo-Strukturen hinzugerechnet wird. Dadurch wird der semantische Wert der ICinfo-Struktur, die

das übergeordnete Konzept darstellt, jedes Mal um 1 erhöht, wenn neue ICinfo-Strukturen hinzukommen, die untergeordnete Konzepte darstellen.

So wie Hyponyme und Meronyme kann ein Wort mehrere Hyperonyme und Holonyme haben. Zum Beispiel können *pet* und *animal* Hyperonyme von *dog* sein, und *car* ist nicht das einzige Holonym von *motor*, denn es gibt zahlreiche anderen Arten von Maschinen, zu denen ein oder mehrere Motoren gehören. Dies trifft auch bei Relationen von Ursache-Wirkung zu. Es wird angenommen, dass die semantische Relevanz eines Wortes mit mehreren Verbindungen zu seinen übergeordneten Wörtern im Text größer als die Relevanz eines Wortes mit weniger Verbindungen ist. Dies kann durch die oben genannte Gewichtungsmethode auch berücksichtigt werden, indem die semantische Gewichtung einer ICinfo-Stuktur bei der neuen Erkennung einer ICinfo-Stuktur automatisch erhöht wird, die das übergeordnete Konzept darstellt.

Es bestehen mehrere Möglichkeiten der Nominalisierung eines Verbs, wie zum Beispiel bei *production* und *producer*, die beide von *produce* abgeleitet sind. Die type-bezogene Berücksichtigung der Anzahl einer lexikalischen Relation wird auch bei der Berechnung dieser Art von Derivation verwendet. Zusammenfassend lässt sich sagen, dass die token-bezogene Berücksichtigung der Anzahl der lexikalischen Relationen nur für die Berechnung der S-Relationen verwendet wird, und die semantische Berechnung aller O-Relationen type-bezogen ist.

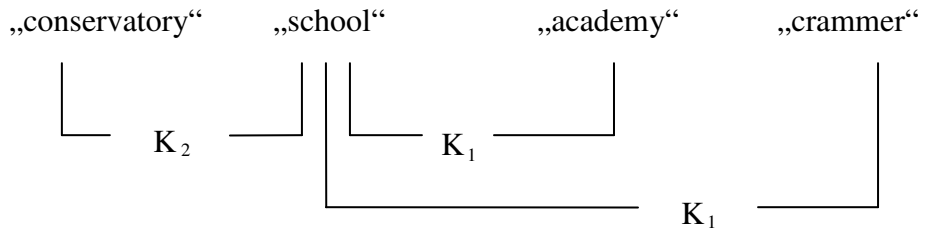
3.3.5 Gewichtsbestimmung der Indexterme

Mit Hilfe der bisher dargestellten Verfahrensweise entsteht die potenzielle lexikalische Kette, die auf mehreren möglichen Konzepten einer Wortform beruhen kann. Ein Lexem kann also durch seine möglichen Belegformen zu mehreren Ketten gehören. Die genaue Ermittlung der passenden Kette unter allen möglichen Ketten ist ein kritisches Problem, da dies einen Rückgriff auf das Weltwissen erfordert, was bei der maschinellen Verarbeitung natürlicher Sprache generell Schwierigkeiten bereitet. Stairmand (1997) und Barzlay/Elhadad (1999:116) versuchen, gültige Ketten von anderen möglichen Ketten zu unter-

scheiden, wobei verschiedene Faktoren wie die Länge, also die Anzahl der Wörter der Kettenoder die Abstände der Wörter berücksichtigt werden. Die Grundidee von Verfahren dieser Art ist, dass Texte normalerweise dazu neigen, möglichst ihren Kontext zu bewahren, die lexikalische Kette mit der größten Wörteranzahl erhält wegen ihres kontextuellen Beitrags somit die höchste Wahrscheinlichkeit für ihre Gültigkeit. Weiter übertragen Stairmand/Black (1997) das Gewicht der ausgewählten lexikalischen Kette auf die ihr zugehörigen Wörter; damit erhalten die einzelnen Wörter dieser Kette, die als Indexterme verwendet werden, dasselbe Gewicht. Durch diesen Prozess wird implizit lexikalische Disambiguierung durchgeführt, indem eine gültige Kette von den Ketten, zu denen das betreffende Wort gehört, selektiert wird und die anderen ausgelassen werden. Dies führt zu der Gefahr, dass eine falsche Bestimmung der Ketten Einfluss auf alle dazugehörigen Konzepte der Ketten hat und dass bei der Bestimmung der längsten Kette diejenigen Konzepte, die zu kürzeren Ketten gehören, nicht in Betracht gezogen werden, obwohl ihnen im Text semantisch gesehen eine wichtige Rolle zukommen kann. Dieses Problem wird noch kritischer, wenn man sich an die in Abschnitt 3.2.3 erwähnte hohe Repräsentationsgranularität des als Wissensbasis eingesetzten WordNet erinnert. Wenn eine lexikalische Kette nur auf einem der möglichen Konzepte gebildet wird und die restlichen Konzepte, die denselben Wortformen zugewiesen werden können, ignoriert werden, dann ist zu erwarten, dass die daraus entstandene lexikalische Kette sehr mager ist und daher semantische Einflüsse von Wörtern, die ähnliche konzeptuelle Informationen darstellen, im Text nicht genug berücksichtigt werden können. Wenn etwa eine lexikalische Kette anhand der zweiten Bedeutungsmöglichkeit von „school“, die als SCHULGEBÄUDE bestimmt ist (vgl. Abschnitt 3.2.3), erstellt wird, werden weitere potenzielle lexikalische Ketten ausgelassen, die auf semantisch ähnlichen Konzepten wie ERZIEHUNGINSTITUT, AUSBILDUNG und MITGLIEDER_VON_ SCHULE, die jeweils die erste, dritte und vierte Belegform von WordNet darstellen, beruhen.

Um Probleme dieser Art möglichst gering zu halten, wird hier statt der Auswahl der längsten Kette das Konzept mit der höchsten Gewichtung der Konzepte, die der gleichen Wortform zugewiesen werden können, ermittelt. Das

Gewicht des Konzepts wird weiter als Gewicht des Indexterms verwendet, der durch die Wortform des Konzepts dargestellt wird. Dies wird ermöglicht, wenn das Analyseverfahren auf Non-Greedy-Algorithmen basiert, durch die keine explizite, streng lexikalische Disambiguierung durchgeführt wird. Bei diesem Verfahren wird die Kette nicht explizit auf der Wortebene gesucht, sondern es sind diejenigen Teilketten von Interesse, die durch die Relationen des Konzepts eines Wortes herausgebildet werden.[50] Wenn die explizite Disambiguierung durch die Auswahl einer lexikalischen Kette erfolgte, dann würde der semantische Beitrag von *conservatory* nicht berücksichtigt, das in Abbildung 3-13 durch SCHULGEBÄUDE mit *school* eine Relation ausmacht.



K 1 : ERZIEHUNGSINSTITUT

K 2 : SCHULGEBÄUDE

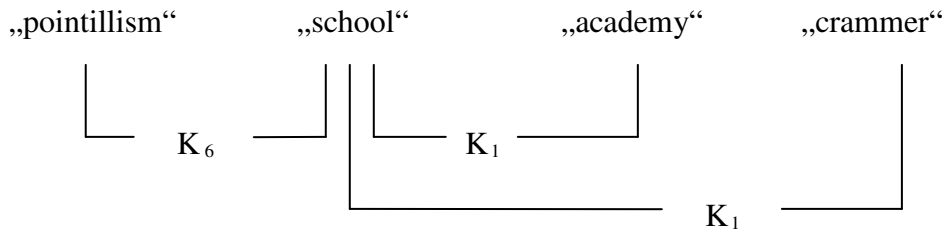
Abbildung 3-13: Zwei mögliche Ketten aus einer Wortform

Da die Auswahl des Konzepts mit dem höchsten semantischen Gewicht einer Wortform die semantischen Gewichte von anderen Konzepten, die zu einer Wortform gehören, vertreten kann, werden folgende zwei Probleme verringert: erstens, dass die längste Kette nicht unbedingt die passende Kette sein muss, und zweitens, dass die Bedeutungen des WordNet zu ausführlich bestimmt sind (zu hohe Granularität der Wissensbasis).

Während Abbildung 3-13 zwei Ketten darstellt, die semantisch sehr ähnlich sind, zeigt Abbildung 3-14 Ketten, bei denen dies nicht der Fall ist. Dass eine Kette von einem Konzept einer Wortform nicht explizit ausgewählt wird, kann dann sinnvoll sein, wenn die längste Kette weniger kontextuelle Relevanz als die kürzere besitzt. Diese Strategie mindert den Nachteil der falschen Auswahl

[50] Der Begriff „Teilkette“ ist immer mit der Perspektive von einem Konzept eines Wortes verbunden. Zum Beispiel findet man in Abbildung 3-13 zwei lexikalische Ketten und vier Teilketten. Die zwei lexikalische Ketten *conservatory – school* und *school – academy – crammer* sind gleichzeitig Teilketten, da sie jeweils aus K_1 und K_2 von *school* gebildet sind. Die Relationen zwischen *school* und *academy* und *school* und *crammer* können weiter als Teilkette betrachtet werden, da sie mit der Perspektive von *academy* und *crammer* anhand K_1 erstellt sind.

einer Kette. Durch die Mitberücksichtigung der anderen Ketten besteht die Möglichkeit, dem semantischen Beitrag anderer Wörter Rechnung zu tragen, die nicht unbedingt zur längsten Kette gehören.



K 1 : ERZIEHUNSINSTITUT
K 6 : KÜNSTLERGRUPPE

Abbildung 3-14: Zwei mögliche Ketten aus einer Wortform

Die entsprechenden Formen der ermittelten Konzepte, die im Text das größte Potenzial zur Darstellung der kognitiven Spuren haben, werden als Darstellungsform des Indexterms zur Gewichtung der Konzepte zugrunde gelegt.

3.4 Vergleich mit dem Vektorraummodell (VRM)

3.4.1 Vektorraummodell

Das Vektorraummodell (VRM) gilt als einer der grundlegenden Ansätze im IR (vgl. Stock 1007:334). Seine Grundidee ist, dass ein Dokument als n-dimensionaler Raum angesehen wird, wobei es aus der n-Zahl der Terme besteht (vgl. Baeza-Yates 1999:27-30, Nohr 2001:26-37, Stock 2007:334-349). Probleme entstehen bei diesem Modell, wenn man versucht, Index- und Retrievalprozess streng zu unterscheiden. Zum Beispiel betrachtet Nohr (2001:26-31) VRM im Rahmen des Indexprozesses, und Baeza-Yates/Ribeiro-Neto (1999:41) betrachten es als Retrieval-Modell. Diese Uneinstimmigkeit liegt vermutlich daran, dass das VRM die damit verbundene Strategie der Termgewichtung impliziert, die streng genommen als Indexierungsprozess betrachtet werden muss, der wiederum mit der Strategie des Retrievals eng verbunden ist. Die Termgewichtung basiert auf der Annahme, dass die Relevanz der Wörter in einem Text durch statistische Methoden ermittelt werden kann, wobei die Häufigkeit des Vorkommens der Wörter im Mittelpunkt steht (vgl. Baeza-Yates/Ribeiro-Neto:24 ff., Nohr 2001:26 ff.). Demnach gilt ein Term für einen Text als ein guter Indexterm, wenn er in diesem sehr häufig, aber selten in anderen Texten vorkommt. Diese beiden Aspekte werden durch die sog. Termfrequenz (*tf*) und die umgekehrte Dokumentfrequenz (*idf*) wie folgt berechnet:

$$tf_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}}$$

$$idf_i = \log \frac{N}{n_i} (+1)$$

Dabei ist $freq_{i,j}$ die Vorkommenshäufigkeit des $Terms_i$ in dem $Dokument_j$, $\max_l freq_{l,j}$ ist die Gesamtzahl aller Terme im $Dokument_j$, N ist die gesamte Zahl der Dokumente, und n_i ist die Zahl der Dokumente, in denen $Term_i$ vorkommt. Folglich wird die Gewichtung des $Term_i$ im $Dokument_j$ wie folgt durch die Multiplikation von *tf* und *idf* berechnet:

$$w_{i,j} = tf_{i,j} * idf_i$$

Bei der Bestimmung der Analyseeinheit ist es üblich, die semantisch unwichtigen Wörter durch das Einsetzen eines Stoppwortliste ausfiltern zu lassen (vgl. Abschnitt 3.2.1). Es werden wie beim kognitiven Ansatz Nomen und Verben für die weitere Arbeit berücksichtigt, damit die beiden Modelle fair verglichen werden können. Für die Geschwindigkeit der Berechnung ist noch zu erwähnen, dass Konstanten, die für die Berechnung des Rankings gebräuchlich sind, wie die Auftrittszahl von Termen in jedem Dokument und in der Dokumentkollektion vorher berechnet und gespeichert werden sollten.

Beim VRM werden Dokumente als Vektoren repräsentiert, wobei die im Dokument vorkommenden Terme als Vektorkomponenten betrachtet werden. Ein Dokument j mit t-Anzahl der Analyseeinheiten, die die Komponenten und die Dimension des Vektors ausmachen, kann mit folgender Formel dargestellt werden:

$$\vec{d_j} = (w_{1,j},\ w_{2,j},\ w_{3,j},\ w_{4j},\ w_{5,j}, \ldots,\ w_{t,j})$$

Die semantische Ähnlichkeit von zwei Dokumenten, die jeweils durch $\vec{a}$ und $\vec{b}$ repräsentiert werden, berechnet man typischerweise durch den Koeffizienten des Kosinus.[51] Dieser beruht auf dem vektoriellen Produkt:

$\vec{a} \bullet \vec{b} = |\vec{a}|\ |\vec{b}| \cos\theta$, wobei θ den geschlossenen Winkel von beiden Vektoren darstellt.

$$\text{sim}(\vec{x}, \vec{y}) = \cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}|\,|\vec{y}|} = \frac{\sum_{i=1}^{m} x_i * y_i}{\sqrt{\sum_{i=1}^{m} x_i^2} * \sqrt{\sum_{i=1}^{m} y_i^2}}$$

Diese Ähnlichkeitsberechnung wendet man nicht nur beim Vergleich der Dokumente im allgemeinen Sinne, sondern auch für den Zweck des Retrievals beim Vergleich zwischen den Dokumenten und einer Benutzeranfrage an, die aus der Anzahl von 0 bis n Suchwörtern besteht und dadurch auch einen Vektorraum bildet.

[51] Zu den anderen Ähnlichkeitsberechnungen der Vektoren s. Maning/Schütze 2001:299, Salton 1989:319.

Wie Voohees (1998:286 f.) und Stairmand (1997:144) zeigen kann man das VRM auch als Methode für die Berechnung der Ähnlichkeit verwenden, wobei sich ihre jeweiligen Verfahren durch die Strategie für die Berechnung des Termgewichts unterscheiden. Beim VRM, das in dieser Arbeit für den Vergleich mit dem kognitiven Ansatz implementiert ist, wird die oben beschriebene Termgewichtsmethode verwendet, die auf der *if* und der *idf* beruht. In dem kognitiven Ansatz wird die Ähnlichkeit zwischen Dokument und Anfragesprache dadurch ermittelt, dass die Summe der Gewichte der mit der Anfragesprache übereinstimmenden Indexterme in einem Dokument berechnet wird.

3.4.2. Zeit

Das Analyseverfahren mit kognitivem Ansatz dauert pro Dokument durchschnittlich ca. 14,7 Sekunden und für 500 Dokumente ca. 122 Minuten.[52] Mit dem VRM-Ansatz hat das ganze Verfahren ca. 19 Minuten gedauert, wobei eine endgültige Gewichtung der Terme wegen der Berechnung von *idf* erst nach der Analyse der gesamten Dokumentkollektion möglich ist. Das Einbeziehen der ganzen Dokumentkollektion bei der Ähnlichkeitsberechnung kann verlangsamend wirken, wenn häufig ein Update der Dokumentkollektion erfolgt.

3.4.3 Rankinganalyse

Beide Textretrieval-Systeme bieten über eine Suchanfrage die Liste der Dokumente an, die nach der Reihenfolge der semantischen Ähnlichkeit mit der Anfrage angeordnet sind. Die Indexierung, die auf der kognitiven Interpretation beruht, wird in der Ergebnistabelle als kognitiver Ansatz genannt. Als erste Anfrage wird *restaurant* eingegeben, dessen Konzept relativ alltäglich ist und von Schank/Abelson (1977:43 f.) exemplarisch für die Erklärung der Konzeptdarstellung verwendet wird. Da die Dokumentenanzahl der Ausgabe überschaubar ist, wird auf die Bestimmung des Schwellenwerts der Dokumente zu Indextermen verzichtet. Als Ergebnis sind folgende 30 Dokumente herausgekommen:

[52] Bei der Evaluierung wird ein Rechner mit einem 1.29-GHz-Intel-Prozessor und mit 256 MB RAM verwendet.

Kognitiver Ansatz				Vektorraummodell(VRM)-Ansatz			
Match-nr.	Doc-Nr.	Freq	Sim	Match-nr.	Doc-Nr.	Freq	Sim
1	240	8	5.5	1	240	8	0.006675
2	31	5	5	2	31	5	0.003084
3	182	3	4	3	182	3	0.002234
4	406	3	4	4	168	5	0.002154
5	188	1	3	5	118	2	0.001764
6	168	5	3	6	406	3	0.001266
7	150	1	3	7	324	2	0.0011
8	116	1	3	8	500	1	0.001042
9	118	2	2.5	9	408	2	0.001014
10	119	2	2.5	10	87	1	0.000804
11	302	1	2	11	17	1	0.000788
12	223	1	2	12	85	1	0.000784
13	405	1	2	13	18	1	0.000733
14	85	1	2	14	223	1	0.000679
15	472	1	2	15	116	1	0.000668
16	42	1	2	16	42	1	0.000662
17	476	1	2	17	472	1	0.000649
18	18	1	2	18	188	1	0.000646
19	324	2	1.5	19	23	1	0.00062
20	408	2	1.5	20	451	1	0.000615
21	434	1	1	21	150	1	0.000586
22	447	1	1	22	53	1	0.000581
23	451	1	1	23	119	2	0.000576
24	500	1	1	24	165	1	0.000554
25	165	1	1	25	302	1	0.000493
26	111	1	1	26	447	1	0.000456
27	87	1	1	27	111	1	0.000456
28	53	1	1	28	434	1	0.00045
29	23	1	1	29	405	1	0.000435
30	17	1	1	30	476	1	0.000421

Tabelle 3-7: Ergebnis 1 (*restaurant*)

Durch den kognitiven Ansatz ist zu erkennen, dass es kein Dokument gibt, in dem das Wort *restaurant* mehr als zwei lexikalische Verbindungen hat. *Restaurant* erweist sich im Korpus als verbindungsunfreundlich im Sinne der lexikalischen Semantik. Da sich beide Ansätze in diesem Fall zum großen Teil auf die Information der Häufigkeit stützen, könnte die Erwartung entstehen, dass nur ein geringer Unterschied bei ihren Rankinglisten zu bemerken ist. Wenn man die ersten sechs Plätze des kognitiven Ansatzes mit den ersten sechs des VRM-Ansatzes vergleicht, so teilen sie sich alle fünf Dokumente mitein-

ander, in denen sich *restaurant* drei bis acht Mal wiederholt. Das Dokumentranking vom 5. bis 30. Platz wird durch die verschiedenen Analysemethoden geprägt. Dies wird besonders für die Dokumente Doc-Nr. 119 und Doc-Nr. 500 ersichtlich, die auf dem 10. bzw. 23. und auf dem 24. bzw. 8. Platz stehen, wobei die unterschiedliche Stärke der semantischen Relevanz von *retaurant* in beiden Dokumenten aus den folgenden Textabschnitten erkennbar ist. Beim kognitiven Ansatz werden die Werte des Indexterms *restaurant* durch die Verbindung der Hyponyme *restaurant – grill* erhöht. Die Verbindung ist akzeptabel, wenn man die Konzeptkette, zu der *restaurant* gehört, und den daraus entstandenen Kontext aus dem folgenden Textabschnitt betrachtet:

T-1 Doc-Nr. 119

[…]
60: decide in the beginning to put your barbecue equipment to work.
61: you can take it with you.
62: a picnic bag , a **grill** , a cooler for soft drinks and beer , and for frozen convenience foods.
63: eat in a **restaurant** or motel mornings and evenings;
64: or just evenings.
65: turn off at any one of the marked picnic areas (gasoline companies have touring service bureaus that issue booklets on national parks to tell you where you have barbecue facilities) and -- with soft drinks cooled from morning loading up , hamburger , buns , an array of relishes , and fresh fruit -- your lunch is 75% cheaper than at **a restaurant** , and 100% more fun.
66: you need a little stove , a coffee pot and a stew pot;
[…]

T-2 Doc-Nr. 500

[…]
64: sometimes he would be up before dawn , clad as a garbage collector and hurling pails into areaways to exasperate us , and thereafter would hurry to the Bronx Zoo to grimace at the lions and press cigar butts against their paws.
65: evenings , he was frequently to be seen at **restaurants** like Enrico & Paglieri's or Peter's Backyard drunkenly donning ladies' hats and singing O Sole Mio.
66: in short , and to borrow an arboreal phrase , slash timber.
67: well , the odious little toad went along chivying animals and humans who couldn't retaliate , and in due course , as was inevitable , overreached himself.
68: one morning , we discovered not only that the pennies were missing from the idol but that a cigarette had been stubbed out in its lap.
[…]

Das Dokument Doc-Nr. 119 ist beim VRM-Ansatz aus verschiedenen Gründen weiter nach hinten gerückt. Ein Grund ist die unterschiedliche Länge der Dokumente, die auf die Berechnung der *tf* Einfluss nimmt, während die *idf* zu allen Termen unabhängig von den Dokumenten konstant bleibt. Ein viel gravierenderer Grund ist jedoch der Unterschied in der Länge des Vektors.[53] Je größer die Länge des Vektors, die bei der Berechnung der Ähnlichkeit den Teil des Nenners bildet, desto kleiner ist der Wert der Ähnlichkeit. Dies wird besonders dann spürbar, wenn sich viele gleiche Wörter in einem Text mehrmals wiederholen. Der Term *restaurant* hat ein Gewicht 0.035168 in Doc-Nr. 119 und 0.02903 in Doc-Nr. 500, aber beide Dokumente haben eine Vektorlänge von 61.0816 bzw. 27.8209, was erheblichen Einfluss auf die Positionierung der beiden Dokumente auf der Rankingliste hat. Der Nachteil des VRM-Ansatzes, dass die semantische Verbindung der Wörter nicht berücksichtigt werden kann, wirkt sich auch bei der Berechnung der Ähnlichkeit kritisch aus. Wenn viele Synonyme von einem Ausdruck mehrmals im Text vorkommen, dann lässt sich dadurch die Ähnlichkeit des Wortes reduzieren, weil die Häufigkeit der Synonyme ein Teil des Nenners ist. Interessant ist, dass nur ein Dokument, nämlich Doc-Nr. 406, zu den ersten zehn der Rankingliste nach kognitivem Ansatz gehört, das als typischer fiktiver Text betrachtet werden kann, während sich unter den ersten zehn nach VRM-Ansatz drei Dokumente dieser Sorte befinden, nämlich Doc-Nr. 406, 408 und 500.

Die Hervorhebung des semantischen Einflusses beim kognitiven Ansatz kann man deutlicher mit folgender Ausgabe der Anfrage *atom* finden:

Kognitiver Ansatz				VRM-Ansatz			
Match-nr.	Doc-Nr.	Freq	Sim	Match-nr.	Doc-Nr.	Freq	Sim
1	101	31	19.5	1	101	31	0.020229
2	298	21	14	2	84	8	0.00771
3	301	1	11	3	298	21	0.006155
4	300	5	6	4	300	5	0.00244
5	84	8	5.5	5	63	3	0.002247
6	88	1	3.5	6	210	1	0.001201

[53] Die Vektorlänge $|\vec{a}|$ darf nicht mit Anzahl der Vektorkomponenten $\vec{a}$ n verwechselt werden.

7	63	3	3	7	20	1	0.000961
8	275	1	3	8	244	1	0.00079
9	366	1	3	9	432	1	0.000735
10	244	1	2	10	366	1	0.000709
11	432	1	2	11	88	1	0.000673
12	20	1	1	12	275	1	0.000494
13	210	1	1	13	301	1	0.000492

Tabelle 3-8: Ergebnis 2 (*atom*)

Bei der Betrachtung der Rankingliste nach dem kognitiven Ansatz sind drei Dokumente, Doc-Nr. 301, 88 und 275, zu betrachten, bei denen das Wort *atom* nur einmal vorkommt und der Ranking-Unterschied zwischen den beiden Ansätzen mehr als vier Plätze ausmacht. Den größten Rankingunterschied zwischen beiden Ansätzen erreicht das Dokument Doc-Nr. 301, in dem *atom* einmal vorkommt. Das Dokument ist beim kognitiven Ansatz auf dem dritten Platz positioniert, beim VRM-Ansatz auf dem letzten. In WordNet sind zwei Konzepte, die mehr oder weniger semantisch ähnlich sind, nämlich 1. als THE SMALLEST COMPONENT OF AN ELEMENT HAVING THE CHEMICAL PROPERTIES OF THE ELEMENT und 2. als A TINY PIECE OF ANYTHING definiert, die möglicherweise „atom“ zugewiesen werden. Die Gewichtung von Dokument Doc-Nr. 301 bei kognitivem Ansatz ist ausschließlich durch das Synonym *particle* von *atom*, das auf der ersten der möglichen Belegformen von „atom“ basiert, erhöht. Der folgende Anfangsabschnitt des Dokuments Doc-Nr. 301 zeigt den Teil des Textes, in dem *atom* und *particle* nahe beieinander vorkommen.

T-3 Doc-Nr. 301

1: the Poynting-Robertson effect (Robertson , 1937;
2: Wyatt and Whipple , 1950) , which is a retardation of the orbital motion of **particles** by the relativistic aberration of the repulsive force of the impinging solar radiation , causes the dust to spiral into the sun in times much shorter than the age of the Earth.
3: the radial velocity varies inversely as the **particle** size -- a 1000-m-diameter *particle* near the orbit of Mars would reach the sun in about 60 million years.
4: Whipple (1955) extends the effects to include the solar-corpuscular-radiation pressure , which increases both the minimum **particle** size and the drag.
5: further , the corpuscular radiation , i.e. , the solar-wind protons , must sputter away the surface **atoms** of the dust and cause a slow diminution in size , with a resultant

increase in both the Poynting-Robertson effect and the ratio of the repulsive force to the gravitational force.
6: the Poynting-Robertson effect causes the semi-major axis of orbits to diminish more rapidly than the semi-minor axis , with a consequent tendency toward circular orbits as the **particles** move toward the sun.
7: also , planetary gravitational attraction increases the dust concentration near the plane of the ecliptic as the sun is approached.
8: at one astronomical unit from the sun (the Earth's distance) the dust orbits are probably nearly circular.
9: if such is the case , the **particles** within a distance of about F of the Earth will have , relative to the Earth , a kinetic energy less than their potential energy and they will be captured into orbits about the Earth.
10: De Jager (1955) has calculated the times required for these **particles** to reach the atmosphere under the influence of the Poynting-Robertson effect , which in this case causes the orbits to become more and more eccentric without changing the semi-major axis.
[…]

Particle in T-3 ist ein in Physik oder Astrophysik domänen-spezifisch verwendeter Begriff, der in WordNet nicht extra definiert ist, deswegen mit *atom* durch sein unspezifisches Konzept in Verbindung gesetzt wird. Dennoch ist es schwer zu bestreiten, dass die Verbindung nicht gültig ist. Wie spezifisch ein Lexikon oder Thesaurus definiert ist, ist eine schwierige Frage, die besser Lexikografen überlassen wird. Es ist hier noch einmal klar geworden, dass der Thesaurus nach seinem Verwendungszweck angemessen konzipiert und definiert werden muss. Der folgende Text, der einen Teil des Dokument Doc-Nr. 88 zeigt, deutet den Vorteil bei klarer Unterscheidung der beiden Konzepte in WordNet an, die *atom* zugewiesen werden können.

T- 4 Doc-Nr. 88

[…]
82: although the fatty protein **molecules** , carried in the blood and partly composed of cholesterol , are water soluble , cholesterol itself is insoluble , and cannot be destroyed by the body.
83: a remarkable substance , says Dr. Keys , quite apart from its tendency to be deposited in the walls of arteries.
84: when thus deposited , Keys says that cholesterol is mainly responsible for the arterial blockages that culminate in heart attacks.
85: explains Keys : as the fatty protein **molecules** travel in the bloodstream , they are deposited in the intima , or inner wall of a coronary artery.
86: the proteins and fats are burned off , and the cholesterol is left behind.
[…]
102: the degree of saturation depends on the number of hydrogen **atoms** on the fat **molecule**.

saturated fats can accommodate no more hydrogens.
103: mono-unsaturated fats have room for two more hydrogens on each **molecule** , and the poly-unsaturated fat molecule has room for at least four hydrogens.
104: the three fats have similar caloric values (about 265 calories per oz.) , but each exerts a radically different influence on blood cholesterol . as a result , although we still make use of this distinction , there is much confusion as to the meaning of the basic terms employed.
[…]

Aus den zwei Verbindungsmöglichkeiten zwischen „atom“ und „molecule“ im Dokument Doc-Nr. 88 wird klar, welche die gültige Verbindung ist. Trotz der Strittigkeit über die Bestimmung der Polysemie scheint die Berücksichtigung einer solchen Verbindung bei der Gewichtung des Indexterms hilfreich zu sein. Dies zeigt sich auch, wenn man das nächste Beispiel anschaut, das einen Teil des Dokuments Doc-Nr. 210 enthält und mit dem VRM-Ansatz auf dem 6. Platz steht, wobei Dokument Doc-Nr. 88 denselben Platz mit dem kognitiven Ansatz belegt.

T- 5 Doc-Nr. 210

[…]
34:the long road that had taken liberals in this country into the social religion of democracy , into a worship of man , led logically to the Marxist dream of a classless society under a Socialist State.
35: and the J existed as the revolutionary experiment in radical socialism , the ultimate exemplar.
36: and by the time the war ended , liberal leadership in this country was spiritually Marxist.
37: we will recall that the still confident liberals of the Truman administration gathered with other Western utopians in San Francisco to set up the legal framework , finally and at last , to rationalize war -- to rationalize want and fear -- out of the world : the United Nations.
38: we of the liberal-led world got all set for peace and rehabilitation.
39: then suddenly we found ourselves in the middle of another fight , an irrational , an indecent , an undeclared and immoral war with our strongest (and some had thought noblest) ally
40: during the next five years the leaders of the Fair Deal reluctantly backed down from the optimistic expectations of the New Deal.
41: during the next five years liberal leaders in the United States sank in the cumulative confusion attendant upon and manifested in a negative policy of Containment -- and the bitterest irony -- enforced and enforceable only by threat of a weapon that we felt the greatest distaste for but could not abandon : the **atom** bomb.
42: in 1952 , it will be remembered , the G.O.P. without positive program campaigned on the popular disillusionment with liberal leadership and won overwhelmingly.

```
43: all of this , I know , is recent history familiar to
you.
44: but I have been at some pains to review it as the drama
of the common man , to point up what happened to him under
Eisenhower's leadership.
45: a perceptive journalist , Sam Lubell , has phrased it
in the title of one of his books as the revolt of the mod-
erates.
46: he opens his discourse , however , with a review of the
Eisenhower inaugural festivities at which a sympathetic
press had assembled its massive talents , all primed to
catch some revelation of the emerging new age.
47: the show was colorful , indeed , exuberant , but the
press for all its assiduity could detect no note of a fate-
ful rendezvous with destiny.
[…]
```

Der Unterschied zwischen den mit *atom* zusammenhängenden Kontexten in den Dokumenten Doc-Nr. 88 und 210 ist gut fassbar. Während das mit *atom* zusammengebundene Wissen in Dokument Doc-Nr. 88 dargestellt ist, ist das in Dokument Doc-Nr. 210 repräsentierte Wissen entfernt vom Konzept ATOM, denn hierin sind weder lexikalische noch grammatische Informationen zu finden, die zwischen dem Wissen über ATOM und anderem Wissen Verbindung stiften. Man kann so sehen, dass zwischen *atom* und dem danach folgenden Wort *bomb* zwar eine kollokative Verbindung besteht, man kann aber den Unterschied zwischen den semantischen Beiträgen von *atom* im den Dokumenten Doc-Nr. 88 und 210 klar erkennen.

Insgesamt ermöglicht das Ergebnis, vorläufig folgende Annahme zu machen: Für die Dokumente, in denen viele Wiederholungen auftreten, ist eine aufwändige semantische Analyse überflüssig; allein die Ermittlung der Wiederholungszahl scheint ausreichend, da der Wert der lexikalischen Verbindung manchmal von der Wiederholung überschattet ist. Dies gilt verstärkt für Dokumente, in denen viele Wörter mit geringem semantischen Verbindungspotenzial vorkommen. Aber wie die Dokumente Doc-Nr. 150, 116 und 119 der Ergebnistabelle für *restaurant* und Doc-Nr. 301 und 88 der Tabelle für *atom* zeigen, unterscheidet sich die Reihenfolge auf der Rankingliste von Dokumenten mit wenigen Wiederholungen bei beiden Ansätzen stark. Die semantische Relevanz eines Konzepts ist durch lexikalische Relationen auch im Text gut hervorgehoben, und dadurch werden die Dokumente, die relativ viele lexikalische Relationen enthalten, beim kognitiven Ansatz nach vorne gerückt. Aus dem Vergleich der Dokumente Doc-Nr. 88 und 210, die auf der jeweiligen

Rankingliste die gleiche Position belegen, ist der Vorteil des kognitiven Ansatzes bei der Ermittlung des Wissens noch einmal sichtbar. Durch lexikalische Relationen wird das mit *atom* verbundene Wissen hervorgehoben. Wenn die Wissensstruktur durch lexikalische Relationen auch im Text deutlich ausgedrückt wird, ist anzunehmen, dass Zusammenhänge zwischen lexikalischen Relationen und der Textsorte bestehen. Um diese Annahmen noch genauer zu festigen, wird *metal* eingegeben, das in 42 Dokumenten auftritt, wovon 29 informative und 13 fiktive Texte sind. Es ergibt sich folgendes Ranking:

Kognitiver Ansatz				VRM-Ansatz			
match-nr.	Doc-Nr.	freq	Sim	Match-nr.	Doc-Nr.	Freq	Sim
1	366	6	13.5	1	366	6	0.00289
2	275	1	7	2	410	4	0.00185
3	296	4	5.5	3	273	5	0.00164
4	273	5	5	4	376	3	0.00147
5	299	2	3.5	5	432	2	0.0001
6	448	2	3.5	6	96	1	0.00097
7	369	2	3.5	7	500	1	0.00092
8	410	4	3.5	8	455	2	0.00086
9	367	2	3	9	296	4	0.00084
10	139	1	3	10	487	1	0.00078
11	101	1	3	11	299	2	0.00073
12	432	2	2.5	12	448	2	0.00069
13	455	2	2.5	13	354	1	0.00065
14	478	1	2.5	14	254	1	0.00065
15	376	3	2	15	478	1	0.00065
16	368	1	2	16	56	1	0.00064
17	462	1	2	17	290	1	0.00062
18	354	1	2	18	70	1	0.00059
19	487	1	2	19	348	1	0.00059
20	500	1	2	20	30	1	0.00056
21	9	1	2	21	279	3	0.00055
22	56	1	2	22	414	1	0.00053
23	96	1	2	23	369	2	0.00051
24	122	1	2	24	387	1	0.00049
25	124	1	2	25	151	1	0.00048
26	145	3	2	26	101	1	0.00044
27	151	1	2	27	9	1	0.00043
28	175	1	2	28	115	1	0.00043
29	279	3	2	29	122	1	0.00041
30	290	3	2	30	111	1	0.00040
31	254	1	1	31	145	1	0.00040
32	385	1	1	32	367	2	0.00039
33	387	1	1	33	385	1	0.00039
34	406	1	1	34	124	1	0.00039

35	407	1	1	35	462	1	0.00038
36	414	1	1	36	406	1	0.00037
37	115	1	1	37	275	1	0.00034
38	111	1	1	38	407	1	0.00030
39	348	1	1	39	368	1	0.00030
40	70	1	1	40	139	1	0.00028
41	341	1	1	41	175	1	0.000267
42	30	1	1	42	341	1	0.00026

Tabelle 3-9: Ergebnis 3 (*metal*)

In der Rankingliste des kohäsionsbasierten IR-Systems gehören acht der ersten zehn Dokumente auf der Rankingliste zu den informativen Texten. Es liegen sogar vier Dokumente, nämlich Doc-Nr. 296, 299, 366 und 369, die informative Texte enthalten, zwischen Doc-Nr. 295 bis 374, die als strenge fachliche Beschreibung zu betrachten sind. Im Gegensatz dazu stehen beim VRM-basierten IR-System nur vier informative Dokumente auf den ersten zehn Plätzen der Rankingliste. Davon sind nur zwei strikt informative Texte. Die Dokumente mit relativ vielen lexikalischen Relationen sind beim kognitiven Ansatz trotz der mangelnden Wiederholung auf der Rankingliste stark nach oben gerückt. Doc-Nr. 410, in dem das Wort *metal* viermal vorkommt, belegt den zweiten Platz auf der Rankingliste des VRM-Ansatzes, im Gegensatz dazu aber den 10. Platz bei einer Analyse nach dem kognitiven Ansatz. Man kann im folgenden entsprechenden Text problemlos erkennen, dass Doc-Nr. 410 trotz des viermaligen Vorkommens von *metal* wenig mit METAL zu tun hat.

T-6 Doc-Nr. 410

```
[…]
26: Mrs. Calhoun has been society editor here for twenty-
five years.
27: the editor says that marriages may be made in heaven ,
but weddings are made in Mrs. Calhoun's columns.
28: she's the one who decides which wedding is to get the
lead space in the Sunday paper and all that.
[…]
74: Carruthers crossed the room to a metal door with an
open grillework in the top half.
75: he pulled it open.
76: now don't shut this door.
77: it won't open from inside.
78: before we built the new jail , we used to keep prison-
ers in here overnight sometimes when the old jail got too
crowded.
79: Hirey treats himself a lot better than we do prisoners.
```

80: they were a sight more comfortable than the ones in the jail with the cold air from Hirey's air conditioner coming through the grille.
81: he walked past the sheriff into a windowless room with shelves full of big , leather-bound volumes from floor to ceiling all around the walls.
82: a **metal** table and four chairs stood in the center.
83: they're all here , back to 1865 , Carruthers told him.
84: it's all right to smoke , but make sure your cigarettes are out before you leave.
85: and , of course , you know not to take clippings.
86: I'll leave the air conditioner on for you , Mr. Ferrell , said Hirey.
87: don't forget to turn it off and close the door good so it'll latch.
88: Hank thanked them and promised to observe the rules.
89: when they had gone , he stood for a minute breathing in the mustiness of old paper and leather which the busily thrumming air conditioner couldn't quite dispel.
90: chapter fourteen
91: in a tour around the stacks , he found that the earliest volumes began on the left and progressed clockwise around the room.
92: an old weakness for burrowing in records rose up to tempt him.
93: it was , indeed , all here -- almost a century.
94: from reconstruction to moon rockets.
95: but he pulled away from the irrelevant old volumes and walked around to the newer ones.
96: last year's volume was at the top a couple of inches below the ceiling.
97: near it was a **metal** ladder on casters attached to the top shelf.
98: he pulled it over , climbed up , and lifted out the big volume , almost losing his balance from the weight of it.
99: he staggered over and dropped it on the table.
100: since Mrs. Calhoun remembered only that the marriage had been in the spring , he started to plod through several months.
101: he tried to turn right to the society page in each one , but interesting stories kept cropping up to distract him.
102: at last he found it in the paper of April 2.
103: it told him little more than Mrs. Calhoun had remembered , stating that it had been a small , modest wedding compared to some of the others.
[…]
117: in the middle of the stock market crash , he heard a slight noise in the outer office.
118: he turned around , saw nothing , and decided it must be a mouse.
119: something else distracted him , yet there was no sound , only tomblike silence.
120: then he knew it was not sound , but lack of it.
121: the air conditioner was no longer running.
122: he jumped up and turned around to see the **metal** door closing.
he smiled.
123: once , when the editor was just out of the hospital from a gallstone operation , Mrs.
124: Calhoun and the mother of the bride went out to his house and fought it out beside his bed.
125: she'd be sure to remember any bride who was vague about background.

[…]

Jeder einzelne Vorkommensfall von *metal* ist jeweils als Attribut von danach folgenden Wörtern zu betrachten, die nicht zur weiteren Entwicklung des Kontextes beitragen. Diese Wiederholungen können als Ganzes den Lesern das Gefühl vermitteln, das mit *metal* verbunden ist, aber keine semantische Signifikanz im strengen Sinne. Zudem ist anzumerken, dass beim kognitiven Ansatz *metal* mit *lead* im Satz 28 verbunden ist und damit eine ungültige Kette entsteht. Diese ungültige Zugehörigkeit und unrichtige Gewichtung könnten vermieden werden, wenn eine genügend hohe Anzahl der Wörter im Text vorkommen würde, die mit dem in diesem Kontext passenden Konzept[54] lexikalische Relationen erstellen.

Als Gegenbeispiel zu Doc-Nr. 410 ist das Dokument Doc-Nr. 275 zu erwähnen, in dem das Wort *metal* nur einmal vorkommt, aber auf der Rankingliste des kognitiven Ansatzes trotzdem auf dem 2. Platz positioniert ist, beim VRM-Ansatz dagegen erst auf dem 37. Platz.

T-7 Doc-Nr. 275

```
1: another recent achievement was the successful develop-
ment of a method for the complete combustion in a bomb
calorimeter of a metal in fluorine when the product is
relatively non-volatile.
2: this work gave a heat of formation of aluminium fluoride
which closely substantiates a value which had been deter-
mined by a less direct method , and raises this property to
15 percent above that accepted a few years ago.
3: similar measurements are being initiated to resolve a
large discrepancy in the heat of formation of another im-
portant combustion product , beryllium fluoride.
4: the development and testing of new apparatus to measure
other properties is nearing completion.
5: in one of these , an exploding-wire device to study sys-
tems thermodynamically up to 6,000 F and 100 atmospheres
pressure , a major goal was achieved.
6: the accuracy of measuring the total electrical energy
entering an exploding wire during a few microseconds was
verified when two independent types of comparison with the
heat energy produced had an uncertainty of less than 2 per-
cent.
7: this agreement is considered very good for such short
time intervals.
8: the method of calibration employs a fixed resistance
element as a calorimeter.
```

[54] In dem Kontext scheint das Konzept von „lead“ angemessen zu sein, das in WordNet als „the intriductory section of a story“ bestimmt ist

```
9: the element is inserted in the discharge circuit in
place of the exploding wire , and the calorimetric heating
of the element is measured with high accuracy.
10: this is used as a reference for comparing the ohmic
heating and the electrical energy obtained from the meas-
ured current through the element and the measured voltage
across the element.
11: a high-speed shutter has been developed in order to
permit photographic observation of any portion of the elec-
trical wire explosion.
12: the shutter consists of two parts : a fast-opening part
and a fast-closing part.
13: using Edgerton's method , the fast-closing action is
obtained from the blackening of a window by exploding a se-
ries of parallel lead wires.
14: the fast-opening of the shutter consists of a piece of
aluminum foil ( approximately F ) placed directly in front
of the camera lens so that no light may pass into the cam-
era.
[…]
72:over a temperature range from 25 to 200 F and at pres-
sures up to 250 J , an overload of 300 J , applied for a
period of one day , results in an uncertainty in the pres-
sure of , at most , one millimeter of mercury.
 transport properties of air.
[…]
104:this apparatus will also be used to measure transition
probabilities of a large number of other elements.
105: a study of the hydrogen line profiles indicates that a
measurement of these profiles can be used to calculate a
temperature for the arc plasma that is reliable to about F
percent.
106: a set of tables containing spectral intensities for
39,000 lines of 70 elements , as observed in a copper ma-
trix in a J arc , was completed and published.
107: studies of the intensity data indicate that they may
be converted to approximate transition probabilities.
108: these data are not of the precision obtainable by the
methods previously mentioned , but the vast number of ap-
proximate values available will be useful in many areas.
109: atomic energy levels.
110:
111: research continues on the very complex spectra of the
rare earth elements.
112: new computer and automation techniques were applied to
these spectra with considerable success.
[…]
```

Insgesamt 13 Wörter kommen im Text vor, die entweder Hyperonyme oder Hyponyme von *metal* sind, wobei das Gewicht von *metal* hauptsächlich durch seine Hyponyme erhöht ist. Insbesondere befinden sich die beiden Hyponyme *aluminium* und *beryllium* direkt in den Sätzen, die auf den Satz folgen, in dem *metal* vorkommt, und als einzelne Beispiele für *metal* dienen. Der semantische Beitrag von *metal* im Text ist durch seine lexikalischen Verbindungen gut erkennbar, wobei das Wort *lead* in diesem Fall in der lexikalischen Kette richtig disambiguiert ist. Der Vergleich mit *restaurant* weist auf den Fall hin, dass ge-

nügend Wiederholungen eines Wortes im Text vorliegen, damit die Anzahl der Wiederholungen für die Gewichtung ausreichend ist. Dem liegt die Annahme zugrunde, dass semantisch wichtige Wörter im Text mehrmals erwähnt werden. Aber die Fortführung kann nicht nur durch Wiederholung, sondern auch durch Synonyme und Hyperonyme realisiert werden. Wenn sich die Textanalyse auf die referenzielle Ebene bezieht, scheint Doc-Nr. 275 trotz des vielmaligen Vorkommens von Hyperonymen und Hyponymen kein gutes Beispiel für ihre Rolle bei der Fortführung derselben Referenten zu sein. Es ist zu erwarten, dass die Lösung des Koreferentenproblems durch lexikalische Relationen besser in denjenigen Texten funktioniert, die ein Objekt oder einen Sachverhalt sachlich beschreiben. Dennoch sind die konzeptuellen Beiträge verschiedener Referenten nicht zu bestreiten. Die Rolle der Hyponyme und der Hyperonyme für die konzeptuellen Beiträge in Doc-Nr. 275 ist sichtbar, was auf den wichtigen Gewinn bei einer Indexierung nach dem kognitiven Ansatz hinweist. Hingegen ist der Nachteil des VRM-Ansatzes problemlos erkennbar, dass nämlich keine Möglichkeit der Berücksichtigung der semantischen Verbindungen besteht, was besonders bei informativen Texten deutlich wird.

Man kann den Schluss ziehen, dass IR mit kognitivem Ansatz im Vergleich mit dem VRM-Ansatz seine Stärke bei der Feststellung und Analyse informativer Texte zeigt. Dies kann darauf beruhen, dass in informativen Texten mehr lexikalische Relationen als in nicht-informativen Texten zu erwarten sind. Deshalb wird diese Stärke bei informativen Texten deutlicher, die durch Wissensstruktur geprägt ist. Dies erinnert an die schema-orientierten Wahrnehmungsvorgänge, anhand deren sich semantische Relationen erklären lassen. Es ist sicherlich kaum zu erwarten, dass, wie bei einem direkten Wahrnehmungsvorgang, lexikalische Relationen wie Meronyme und Hyponyme in einem Text jedes Mal auf eine unmittelbare semantische Verbindung zwischen einem Konzept und seinen Konzeptmerkmalen hinweisen. Aber es ist nicht zu bestreiten, dass die Hyperonyme und Holonyme auf der Textmakroebene die Domänen herstellen, die auf ihre Hyponyme und Meronyme semantischen Einfluss nehmen. Umgekehrt befestigen alle möglichen Vorfälle oder Ereignisse in einer Domäne deren Realisierung. Dies wird deutlicher bei informativen Texten, die zu einem großen Teil menschliche konzeptuelle Struktur wiedergeben. Die lexikali-

sche Kette unterstützt einerseits die konzeptuelle Erstellung des Kontextes und zeigt andererseits den Zusammenhang mit dem pragmatischen Textcharakter, der sich bei der Klassifizierung der Textsorte auswirkt. Dies führt zu der Annahme, dass ein Zusammenhang zwischen der Zahl und der Art von lexikalischen Relationen und Textsorte besteht. Weitere eingehende Untersuchungen über das genaue Verhältnis sind wünschenswert, weil sie für die automatische Textklassifikation eingesetzt werden könnten.

4. Schlusswort

Es wurde gezeigt, dass eine kohäsions-basierte Indexierung im Vergleich zum VRM normalerweise zu höherer Präzision beiträgt. Das größte Problem liegt in der Diskrepanz zwischen dem im WordNet repräsentierten semantischen Wissen und dem des Benutzers. Der Indexierungsprozess ist abhängig von dem semantischen Wissen in WordNet, das sehr fein bestimmt ist, aber diese Feinbestimmung und die daraus resultierenden aufwändigen Berechnungen scheinen teilweise überflüssig für die normale Nutzung zu sein, bei der kein sehr spezifisches Wissen nötig ist. Die Ausnutzung der Wissensbasis hängt mit dem Übereinstimmungsgrad mit der Ausführlichkeit seiner Wissensrepräsentation und den Informationsbedürfnissen des Benutzers zusammen. Wie in Abschnitt 3.3.1.1 gezeigt wurde, ist bei der Repräsentationsgranularität die Betrachtung der Transitivität problematisch. Einige Faktoren wurden erwähnt, die beim Einsetzen der Pfadlänge berücksichtigt werden müssen: die Ausdrucksweise der Textproduzenten für die Realisierung der Koreferenz, die Repräsentationsgranularität des lexikalischen Systems, die Länge der Hierarchien, in denen sich die Wörter befinden, und die Höhe der einzelnen Wörter in den Hierarchien. Über diese Faktoren sind weitere Forschungen wünschenswert. Aus den bisherigen Überlegungen lässt sich schließen, dass bei der Modellierung und dem Einsetzen der Wissensbasis die Interessen und Informationsbedürfnisse der Benutzer berücksichtigt werden müssen und man von der Nutzung des IRS beim domänenspezifischen Wissen am besten profitieren kann.

Wie Bazilay/Elhadad (1999:118) andeuten, ist die Berücksichtigung der Anaphora bei der Kohäsionsanalyse ein kritisches Problem. Die Fortführung eines Konzepts oder eines Referenten kann nicht nur durch Wiederholung der entsprechenden Vollwörter, sondern auch durch andere Anaphernmittel[55] wie Pronomen oder sogar durch Ellipsen realisiert werden. Die semantische Gewichtung der Indexterme kann besonders dann nicht fair berechnet werden, wenn Textproduzenten die Verwendung von Pronomina bevorzugen. Die Forschung bestätigt, dass die Anaphernresolution zu einer besseren Retrievalqualität beiträgt (Stock 2007:297 f.). Der Nachteil ist, dass die Anaphernresolution

[55] Zur Klassifizierung der Anaphora s. Mitkov 2002:27.

komplexere und aufwändigere Prozesse benötigt, die sich auf fast alle linguistischen Analysenebenen wie die morphologische, semantische und syntaktische Ebene beziehen (Mitkov 2002). Ein Beispiel hierfür ist die Phrasenerkennung. Dies kann als allgemeines Problem gesehen werden, wenn computerlinguistische Methoden für die Inhaltsanalyse verwendet werden. Wie in Abschnitt 2.1 gezeigt wurde, können Hyponyme durch entsprechende Hyperonyme mit passendem adjektivistischem Modifikator dargestellt werden. Solche einfachen Phrasen sind besonders im Englischen schwer von den Komposita zu unterscheiden, insofern sie beide zu einer semantischen Einheit aus mehreren Wörtern zusammengesetzt sind, wobei zwischen den Wörtern Leerstellen vorkommen. Ob die Komposita im Hinblick auf die Qualität des IRS als Komposita oder als einzelnes Wort erkannt werden sollen, ist eine schwierige Frage. Aber besonders für die lexikalische Kohäsion ist die Erkennung der Komposita äußerst wünschenswert, weil die auf diesem Wege ermittelten Kohäsionsketten noch genauer die semantischen Informationen von Texten darstellen können, indem die Komposita als angemessene semantische Einheit innerhalb der betreffenden Texte berücksichtigt werden.
Im Hinblick auf die Bestimmung der Analyseeinheiten besteht auch bei Eigennamen ein ähnliches Problem, sofern sie durch Leerstellen in mehrere Teile zerlegt werden können, wie *Senator Barack Obama*, *Calvin Klein* etc. Semantisch betrachtet sind Eigennamen Klassen, die nur ein Klassenmitglied enthalten, weswegen ohne andere sprachliche Mittel direkt darauf referiert werden kann (vgl. Bußmann 2002:185).[56] Man kann damit rechnen, dass besonders in einem domänenspezifischen IR die Retrievalqualität erhöht wird, wenn Eigennamen als Instanzen von anderen Wörtern berücksichtigt werden können, zum Beispiel *Bob Dylan* als *Singer* und *Barack Obama* als *Senator*.

Insgesamt kann man sehen, dass der Versuch, eine bessere Retrievalqualität durch computerlinguistische Methoden zu erreichen, zur vollständigen Sprachanalyse hinführt, die mit Textverstehen in Verbindung gebracht werden kann. Dies heißt nicht unbedingt, dass alle automatischen Textanalysen letztlich nach vollständigem Textverstehen streben müssen. Das Textverstehen erfordert hierbei eine sehr komplexe Wissensbasis sowie einen hohen Analyseaufwand

[56] Zu dem Problem der Erkennung von Eigennamen s. Stock 2007:256 f.

und funktioniert normalerweise mit einer starken Kontextbeschränkung besser. Die Wahl der computerlinguistischen Methode und die Granularität der Textanalyse in IR sollten sich nach verschiedenen Faktoren richten. Zu letzteren gehören zum Beispiel die zur Verfügung stehende Wissensbasis sowie die Frage, ob der Vorgang als allgemeine oder als domänenspezifische Suchfunktion konzipiert wird und in welcher Sprache die Dokumentenkollektion vorliegt.

5. Literaturverzeichnis

Wörterbücher

Althaus, H. P./Henne, H./Wiegand, H. E. (eds.) (1980) Lexikon der Germanistischen Linguistik. Tübingen: Niemeyer.

Asher, R. E./Simpson, J. M. Y. (eds.) (1993) The Encyclopedia of Language and Linguistics, Bd. I – X. Oxford et al.: Pergamon.

Bußmann, H. (ed.) (2002) Lexikon der Sprachwissenschaft. Stuttgart: Kröner.

Häckler, H. O./Stapf, K.-H. (eds.) (2004) Dorsch Psychologisches Wörterbuch. Bern et al.: Huber.

Lewandowski, T. (1994) Linguistisches Wörterbuch, Bd. I–III. Heidelberg/Wiesbaden: UTB.

Prechtl, P./Burkard, F.-P. (eds.) (1999) Metzler-Philosophie-Lexikon. Stuttgart/Weimar: Metzler.

Sinclair, J. (ed.) (1990) Collins Cobuild English Grammar. London/Glasgow: Collins Cobuild.

Zilahi-Szabo, M. G. (ed.) (1995) Kleines Lexikon der Informatik und Wirtschaftsinformatik. München/Wien: Oldenbourg.

Sonstige Literatur

Abecker, A./v. Elst, L. (2004) Ontologies for Knowledge Management. In: Staab, S./Studer, R. (eds.) (2004). Handbook of Ontologes. Berlin et al.: Springer, 435-454.

Abelson, R. P. (1976) Script Processing in attitude Formation and Decision Making. In: Carroll, J. S./Payne, J. W. (eds.) (1976) Cognitive and Social Behavior. Hillsdale: Erlbaum, 33-46.

Aebli, H. (2001) Denken, das Ordnen des Tuns: Kognitive Aspekte der Handlungstheorie, Bd. I. Stuttgart: Klett-Cotta.

Aebli, H. (1994) Denken: das Ordnen des Tuns: Denkprozesse, Bd. II. Stuttgart: Klett-Cotta.

Aebli, H. (1988) Begriffliches Denken. In: Mandl, H./Spada, H. (eds.) (1988). Wissenspsychologie. München/Weinhein: Psychologie Verlag, 227-246.

Aitchison, J. (1985) Cognitive Clouds and Semantic Shadows. In: Language and Communication 5, 69-93.

Aitchison, J. (1994) Words in the Minds: An Introduction to the Mental Lexicon. Oxford: Blackwell (dt. Übers. Wörter im Kopf: Eine Einführung in das mentalen Lexikon. Tübingen: Niemeyer, 1997).

Al-Halimi, R./Kazman, R. (1998) Temporal Indexing through Lexical Chaining. In: Fellbaum, C. (ed.) (1998) Wordnet: An Electronic Lexical Database. Cambridge/London: MIT Press, 333-352.

Ammann, H. (1928) Die menschliche Rede: Sprachphilosophische Untersuchungen, Teil 2. Darmstadt: Lahr.

Anderson, J. R. (1983) The Architecture of Cognition. Cambridge et al.: Harvard University Press.

Anderson, J. R. (1995) Learning and Memory: An integrated Approach. New York et al.: John Wiley & Sohns, INC.

Anderson, J. R. (1995a) Cognitive Psychology and its Implies. New York: Freeman (dt. Übers. Kognitive Psychologie: Eine Einführung. Heidelberg: Spektrum, 1996).

Anderson, S. R. (1985) Typological distinction in word formation. In: Shopen, T. (ed.) (1985) Language typology and syntactic description, Bd. III: Grammatical categories and the lexicon. Cambridge: Cambridge University Press, 3-56.

Armstrong, S.L. /Gleitmann, L./ Gleitman, R. (1983) What Some Concept might not be. In: Cognition 13, 263-308.

Arnold, H. L./ Sinemus, V. (eds.) (1973) Grundzüge der Literatur- und Sprachwissenschaft, Bd. I. München: DTV.

Atkins, B. T. S./Zampolli, A. (1994) Computational Approaches to the Lexicon. Oxford: Oxford University Press.

Bach, E./Harms, R. T. (eds.) (1968) Universals in linguistic Theory. New York: Holt, Rinehart and Winston, Inc.

Baddeley, A. D. (1990) Human Memory: Theory and Practice. Hove: Erlbaum.

Baeza-Yates, R./Ribero-Neto, B. (1999) Modern Information Retrieval. Essex: Addison Wesley.

Ballstaedt, S.-P./Mandl, H./Schnotz, W./Tergan, S.-O. (1981) Texte verstehen, Texte gestalten. München et al.: Urban & Schwarzberg.

Bara, B. G./Guida, G. (eds.) (1984) Computational Models of Natural Language Processing. Amsterdam/New York: North-Holland.

Bartlett, F. C. (1932) Remmembering: A Study in Experimental and Social Psychology. Cambridge: Cambridge University Press.

Barzilay, R./Elhadad, M. (1999) Using Lexical Chains for Text Summarization. In: Mani, I./Mabury, M. T. (eds.) (1999) Advances in Automatic Text Summarization. Cambridge/London: MIT Press, 111-122.

Bate, A./McNew, S./MacWhinney, B./Devesocvi, A./Smith, S. (1982) Functional Constraints on Sentence Processing: A Crosslinguistic study. In: Cognition 11, 245- 299.

Bátori, I., S./Lenders, W./Putschke, W. (eds.) (1989) Computational Linguistics/Computer-Linguistik. An International Handbook on Computer Oriented Language Research and Application. Berlin/New York: de Gruyter.

Bauer, L. (1983) English Word-Formation. Cambridge: Cambridge University Press.

Bauer, S. Faschinger, M. (2003) Information Search and Retireval. In: http://www.iicm.tugraz.ac.at/cguetl/education/isr/vo/inhalte/block02/Zusammenfassung.html (20.09. 2006).

Baum, R./ Böckle, K./ Hausmann, F. J./Lebsanft, F. (eds.) (1994) Lingua et Traditio: Geschichte der Sprachwissenschaft und der neueren Philologien. Tübingen: Narr.

Bazell, C. E./Catford, J. C./Halliday, M. A. K./ Robins, R. H. (eds.) (1966) In Memory of J. R. Firth. London: Longman.

Bean, C. A./Green, R. (eds.) (2001) Relationships in the Organization of knowledge. Dordrecht et al.: Kluwer Academic Press.

de Beaudegard, R.-A./Dressler, W. U. (1981) Einführung in die Textlinguistik. Tübingen: Niemeyer.

Beckenkamp, M. (1995) Wissenspsychologie: Zur Methodologie kognitionswissenschaftlicher Ansätze. Heidelberg: Roland Asager.

Behrens, L./ Sasse H.-J. (1997) Lexical Typology: A Programmatic Sketch. Köln: Universität Köln.

Bekavac, B. (2001) Information Retrieval. In: http://www.inf-wiss.uni-konstanz.de/CURR/winter0102/IR/ir_script_ws01.pdf (10.10. 2006).

Belke, H. (1973) Gebrauchtext. In: Arnold, H. L./Sinemus, V. (eds.) (1973) Grundzüge der Literatur- und Sprachwissenschaft, Bd. I. München: DTV, 320-341.

Bertram, J. (2005) Einführung in die inhaltliche Erschließung: Grundlage-Methode-Instrumente. Würzburg: Ergon.

Beyer, R. (2003) Verstehen von Diskursen. In: Rickheit, G./Herrmann, Th./Deutsch, W. (eds.) (2003) Psycholinguistik: Ein internationales Handbuch. Berlin/New York: de Gruyter, 532-544.

Bierwisch, M. (1983) Semantische und konzeptuelle Repräsentation lexikalischer Einheiten. In: Ružička, R./Motsch, W. (eds.) (1983) Untersuchungen zur Semantik. Berlin: Akademie Verlag (= studia grammatica XXII), 61-99.

Bierwisch, M. (1991) Vergangenheit und Zukunft der kognitiven Linguistik. In Linguistische Studien 209, 1-6.

Bierwisch, M. /Lang, E. (1987) Grammatische und konzeptuelle Aspekte von Dimensionsadjektiven. Berlin: Akademie Verlag.

Blank, A. (2001) Einführung in die lexikalische Semantik: für Romanisten. Tübingen: Niemeyer.

Bloom, P. (2000) How Children Learn the Meanings of Words. Cambridge et al.: MIT Press.

Bobrow, A. M./Collins, A. M. (eds.) (1975) Representation and Understanding. New York: Academic Press.

Bobrow, A. M./Norman D. A. (1975) Some Principles of Memory Schemata. In: Bobrow, A. M./Collins, A. M. (eds.) (1975) Representation and Understanding. New York: Academic Press, 313-149.

Bolinger, D. L. (1969) Categories, Features, Attributes, In: Brno Studies in English 8, 38-41.

Breuer, D. (1974) Einführung in die pragmatische Texttheorie. München: Fink.

Brinker, K. (2000) Textstrukturanalyse. In: Brinker, K./Antos, G./Heinemann, W./Sager, S. F. (eds.) (2000) Text- und Gesprächslinguistik, Bd. I. Berlin/New York: de Gruyter, 164-175.

Brinker, K. (2000a) Textfunktionale Analyse. In: Brinker, K./Antos, G./Heinemann, W./Sager, S. F. (eds.) (2000) Text- und Gesprächslinguistik, Bd. I. Berlin/New York: de Gruyter, 175-186.

Brinker, K. (2005) Linguistische Textanalyse: Eine Einführung in Grundbegriffe und Methoden. Berlin: Schmidt.

Brinker, K./Antos, G./Heinemann, W./Sager, S. F. (eds.) (2000) Text- und Gesprächslinguistik, Bd. I. Berlin/New York: de Gruyter.

Brown, C. H. (2002) Paradigmatic relation of inclusion and identity I: Hyponymy. In: Cruse et al. 2002, 472-480.

Brown, C. H. (2002a) Paradigmatitic relations of inclusion and identity II: Meronymy. In: Cruse, D. A./Hundsnurscher, F./Job, M./Lutzeier, P. R. (eds.) (2002) Lexikologie: Ein internationales Handbuch zur Natur und Struktur von Wörter und Wortschätzen, Bd. 1. Berlin/New York: de Gruyter, 480-485.

Buckland, M. (1991) Information and Information Systems. New York et al.: Greenwood Press.

Busse, D. (1992) Textinterpretation, Sprachtheoretische Grundlagen einer explikativen Semantik. Opladen: Westdeutscher Verlag.

Bühler, K. (1934) Sprachtheorie. Jena: Fischer.

Bühler, K. (1966) Über Gedanken. In: Graumann, C. F. (ed.) (1966) Denken. Köln/Berlin: Kiepenheuer/Witsch, 60-73.

Carey, S./Bartlett, E. (1978) Acquiring a Single New Word. In: Papers and Reports on Child Language Development 15, 17-29.

Carnap, R. (1952) Meaning Postulate. In: Philosophical Studies 3, 65-73.

Carpenter, P. A./Just, M. A. (1977) Reading Comprehension as Eyes See It. In: Just, M. A./Carpenter, P. A. (eds.) (1977) Cognitive Processes in Comprehension. Hillsdale: Erlbaum, 109-139.

Carroll, J. S./Payne, J. W. (eds.) (1976) Cognitive and Social Behavior. Hillsdale: Erlbaum.

Chafe, W. L. (1970) Meaning and the Structure of Language. Chicago/London: The University of Chicago Press.

Chaffin, C. C./Herrmann, D. J. (1988) The Nature of Semantic Relations: A Comparison of two Approaches. In: Evens, M. W. (ed.) (1988) Relational Models of the Lexicon. Cambridge: Cambridge University Press, 289-334.

Chaffin, R. (1992) The Concept of a Semantic Relation. In: Kittay, E. F. /Lehrer, A. (eds.) (1992) Frames, Fields, and Contrast, Hillsdale, N.J.: Erlbaum, 253-288.

Chowdhury, G. G. (1999) Introduction to Modern Information Retrieval. London: Library Association Publishing.

Christmann, U. (1989) Modelle der Textverarbeitung: Textbeschreibung als Textverstehen. Münster: Aschendorff.

Christmann, U. (2000) Aspekte der Textarbeitungsforschung. In: Brinker, K./Antos, G./Heinemann, W./Sager, S. F. (eds.) (2000) Text- und Gesprächslinguistik, Bd. I. Berlin/New York: de Gruyter, 113-122.

Chu, H. (2003) Information Representation and Retrieval in the Digital Age. Medford/New York: Information Today.

Clark, E. V. (1992) Convensionality and Contrast. In: Kittay, E. F. /Lehrer, A. (eds.) (1992) Frames, Fields, and Contrast, Hillsdale, N.J.: Erlbaum. 171-188.

Clark, E. V. (1993) The Lexicon in Acquisition. Cambridge: Cambridge University Press.

Cleveland, D. B./Cleveland, A. D. (2001) Introduction to Indexing and Abstracting. Englewood, Colorado: Libraries Unlimited, Inc.

Cole, P. (ed.) (1978) Syntax and Semantics: Pragmatics, Bd. IX. New York: Academic Press.

Cole, P. (ed.) (1981) Radical Pragmatics. New York: Academic Press.

Cole, P./Morgan, J. L. (eds.) (1975) Syntax and Semantics: Speech Acts, Bd. III. New York: Academic Press.

Cole, P./Sadock, J. M. (eds.) (1977) Syntax and Semantics: Grammatical Relations, Bd. V III. New York: Academic Press.

Cole, R. W. (ed.) (1977) Current Issues in Linguistic Theory. Bloomington et al.: Indiana University Press.

Collins, A. M./Loftus, E. F. (1975) A Spreading-Activation Theory of Semantic Processing. In: Psychological Review 82, 407-428.

Collins, A. M./Quillian, M. R. (1969) Retrieval Time from Semantic Mermory. In: Journal of Verbal Learning and Verbal Behavior, 240-247.

Combrie, B. (ed.) (1981) Language Universals and Sytactic Typology: Syntax and Morphology. Chicago: University of Chicago Press.

Combrie, B. (1981a) Causative Verb Formation and Other Verb-deriving Morphology. In: Combrie 1981, 309-348.

Conrad, C. (1972) Cognitive Economy in Semantic Memory. In: Journal of Experimental Psychology 92, 149-154.

Copeland, J. E. (ed.) (1984) New Directions in Linguistics and Semiotics. Houston, Texas: Rice University.

Cormen, T. H./Leiserson, C. E./Rivest, R. L/Stein, C. (2001) Introduction to Algorithms. Cambridge: MIT Press.

Coseriu, E. (1967) lexikalische Solidarität. In: Geckeler, H.(ed.) (1978) Strukturelle Bedeutungslehre. Darmstadt: Wissenschaftliche Buchgesellschaft, 239-253.

Coseriu, E. (1978) Probleme der strukturellen Semantik. Tübingen: Narr.

Cotton, J. W./Klatzky, R. L. (eds.) (1978) Semantic Factors in Cognition. Hillsdale, N. J.: Erlbaum.

Craig, C. (ed.) (1986) Noun Classes and Categorization. Amsterdam: Benjamins.

Croft, W./Cruse, D. A. (2004) Cognitive Linguistics. Cambridge: Cambridge University Press.

Cruse, D. A. (1986) Lexical Semantics. Cambridge: Cambridge University Press.

Cruse, D. A. (2000) Meaning in Language: An Introduction to Semantic and Pragmatics. Oxford: Oxford University Press.

Cruse, D. A. (2002) Dimensions of Meaning II: Descriptive Meaning. In: Cruse, D. A./Hundsnurscher, F./Job, M./Lutzeier, P. Rolf, (eds.) (2002) Lexikologie: Ein internationales Handbuch zur Natur und Struktur von Wörter und Wortschätzen, Bd. I. Berlin/New York: de Gruyter, 350-355.

Cruse, D. A. (2002a) Descriptive Models for Sense Relations II: Cognitive Semantics. In: In Cruse, D. A./Hundsnurscher, F./Job, M./Lutzeier, P. R. (eds.) (2002) Lexikologie: Ein interna-

tionales Handbuch zur Natur und Struktur von Wörter und Wortschätzen, Bd. I. Berlin/New York: de Gruyter, 542-549.

Cruse, D. A. (2002b) Paradigmatitic relations of inclusion and identity III: Synonymy. In: Cruse, D. A./Hundsnurscher, F./Job, M./Lutzeier, P. R. (eds.) (2002) Lexikologie: Ein internationales Handbuch zur Natur und Struktur von Wörter und Wortschätzen, Bd. I. Berlin/New York: de Gruyter, 485-492.

Cruse, D. A. (2002c) Hyponymy and Its Varieties. In: Green, R./Bean, C. A./Myaeng, S. H. (eds.) (2002) The Semantics of Relationships: An Interdisciplinary Perspective. Dordrecht et al.: Kluwer Academic Publishers, 3-21.

Cruse, D. A./Hundsnurscher, F./Job, M./Lutzeier, P. R. (eds.) (2002) Lexikologie: Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wortschätzen, Bd. I. Berlin/New York: de Gruyter.

Daneš, F. (1964) A Three Level Approach to Syntax, In: Travaux Linguistiques de Prague 1, 226-240.

Daneš, F. (1970) Zur linguistischen Analyse der Textstruktur. In: Folia Linguistica 4, 72-78.

Daneš, F. (ed.) (1974) Papers on Functional Sentence Perspective. The Hague: Mouton.

Daneš, F./Viehweger, D. (eds.) (1976) Probleme der Textgrammatik. Berlin: Akademie Verlag(= studia grammatica XI).

Daneš, F./Viehweger, D. (eds.) (1977) Probleme der Textgrammatik 2. Berlin: Akademie Verlag (= studia grammatica XVIII).

Deese, J. (1965) The Structure of Associations in Language and Thought. Baltimore, Md.: Johns Hopkins University Press.

Diestel, R. (2006) Graph Theory. Heidelberg: Springer.

Dietze, J. (1994) Texterschließung: Lexikalische Semantik und Wissensrepräsentation. München et al: Sauer.

van Dijk, T. A. (1972) Foundation for Typology of Texts. In: Semiotica 6, 297-323.

van Dijk, T. A. (1977) Text and Context: Explorations in the Semantics and Pragmatics of Discourse. London: Longman.

van Dijk, T. A. (1980) Textwissenschaft. München: DTV.

van Dijk, T. A. (1980b) Macrostructures. Hillsdale, N.J.: Erlbaum.

van Dijk, T. A./Kintsch, W. (1978) Toward a Model of Text Comprehension and Production. In: Psychological Review 85, 363-394.

van Dijk, T. A./ Kintsch, W. (1983) Strategy of Discourse Comprehension. New York et al.: Academic Press.

Dik, S. C. (1978) Functional Grammar. Amsterdam: North-Holland.

Dimter, M. (1981) Textklassenkonzepte heutiger Alltagssprache: Kommunikationssituation, Textfunktion und Textinhalt als Kategorien alltagssprachlicher Textklassifikation. Tübingen: Niemeyer.

Dressler, W. (1972) Einführung in die Textlinguistik. Tübingen: Niemeyer.

Dressler, W. (ed.) (1977) Current Trends in Textlinguistics. Berlin/New York: de Gruyter.

Dressler, W. (ed.) (1978) Textlinguistik. Darmstadt: Wissenschaftliche Buchgesellschaft.

Dressler, W. (1981) Einführung in die Textlinguistik. Tübingen: Niemeyer.

Drosdowski, G./Henne, H./Wiegand H. E. (eds.) (1977) Nachdenken über Wörterbücher. Mannheim: Bibliographisches Institut.

Edmonds, P./Hirst, G. (2002) Near-Synonymy and Lexical Choice. In: Computational Linguistics 28, 105-144.

Ehlich, K. (1984) Zum Textbegriff. In: Rothkegel, A./Sandig, B. (eds.) (1984) Text-Textsorte-Semantik. Hamburg: Buske, 9-25.

Eigler, G./Jechle, T./Merziger, G./Winter A. (1990) Wissen und Textproduzieren. Tübingen: Narr.

Eikmeyer, H.-J./Kindt, W./Laubenstein, U./Lisken, S. Riesler, H./Schade, U. (1995) Coherence Regained. In: Rickheit, G./Habel, C. (eds.) (1995) Focus ans Coherence in Discourse Processing. Berlin/New York: de Gruyter, 115-142.

Eikmeyer, H.-J./Riesler, H. (eds.) (1981) Words, Words and Contexts: New Approaches in Word Semantics. Berlin/New York: de Gruyter.

Eikmeyer, H.-J./ Riesler, H. (1981a) Meanings, Intensions, and Stereotypes: A New Approach to Linguistic Semantics, In: Eikmeyer, H.-J./Riesler, H. (eds.) (1981) Words, Words and Contexts: New Approaches in Word Semantics. Berlin/New York: de Gruyter, 133-150.

Endres-Niggemeyer, B./Krause, J. (eds.) (1985) Sprachverarbeitung in Information und Dokumentation, Jahrestagung der GLDV. Berlin et al.: Springer.

Engelkamp, J. (1976) Satz und Bedeutung. Stuttgart: Kohlhammer.

Engelkamp, J. (1983) Word Meaning and Word Recognition. In: Seiler, T. B./Wannenmacher, W. (eds.) (1983) Concept Development and the Development of Word Meaning. Berlin/Heidelberg: Springer, 17-33.

Engelkamp, J. (ed.) (1984) Psychologische Aspekte des Verstehens. Berlin/Heidelberg: Springer (= Lehr und Forschungstexte Psychologie 10).

Engelkamp, J. (1984a) Sprachverstehen als Informationsverarbeitung. In: Engelkamp, J. (ed.) (1984) Psychologische Aspekte des Verstehens. Berlin/Heidelberg: Springer (= Lehr und Forschungstexte Psychologie 10), 31-53.

Engelkamp, J. (1985) Die Repräsentation der Wortbedeutung. In: Schwarze, C./Wunderlich, D. (eds.) (1985) Handbuch der Lexikologie. Königstein: Athnäum, 292-313.

Engelkamp, J. /Pechmann, T. (1988) Kritische Anmerkung zum Begriff der mentalen Repräsentation. In: Sprach und Kognition 7, 2-11.

Engelkamp, J. /Zimmermann, H., (1983) Foci of Attention in Comprehension and Production of Sentence. In: Rickheit, G., /Bock, M. (eds.) (1983) Psycholinguistic Studies in Language Processing. Berlin/New York: de Gruyter, 119-133.

Erickson, T. D. /Mattson, M. E. (1981) From Word to Meaning: A Semantic Illusion. In: Journal of Verbal Learning and Verbal Behavior 20, 540-551.

Estes, W. K. (1994) Classification and Cognition. New York/Oxford: Oxford University Press.

Evens, M. W. (ed.) (1988) Relational Models of the Lexicon. Cambridge: Cambridge University Press.

Faloutsos, C. (1992) Signature Files. In: Frakes, W., B./ Baeza-Yates, R. (eds.) (1992) Information Retrieval and Data Structures. Upper Saddle River, N. J.: Prentice Hall, 44-65.

Favre-Bulle, B. (2001) Information und Zusammenhang: Informationsfluss in Prozessen der Wahrnehmung, des Denkens und der Kommunikation. Wien/New York: Springer.

Feilke, H. (2000) Die pragmatische Wende in der Linguistik. In: Brinker, K./Antos, G./Heinemann, W./Sager, S. F. (eds.) (2000) Text- und Gesprächslinguistik, Bd. I. Berlin/New York: de Gruyter, 64-82.

Fellbaum, C. (ed.) (1998) Wordnet: An Electronic Lexical Database. Cambridge/London: MIT Press.

Fellbaum, C. (1998a) Introduction. In: Fellbaum, C. (ed.) (1998) Wordnet: An Electronic Lexical Database. Cambridge/London: MIT Press, 1-19.

Fellbaum, C. 1998b, A Semantic Network of English Verbs. In: Fellbaum, C. (ed.) (1998) Wordnet: An Electronic Lexical Database. Cambridge/London: MIT Press, 69-104.

Figge, U. L. (1994) Sprache dient zum Ausdruck von Gedanken. : Zur Geschichte der Formulierung. In: Baum, R./ Böckle, K./ Hausmann, F. J./Lebsanft, F. (eds.) (1994) Lingua et Traditio: Geschichte der Sprachwissenschaft und der neueren Philologien. Tübingen: Narr, 561-665.

Figge, U. L. (2000) Die kognitive Wende in der Textlinguistik. In: Brinker, K./Antos, G./Heinemann, W./Sager, S. F. (eds.) (2000) Text- und Gesprächslinguistik, Bd. I. Berlin/New York: de Gruyter, 96-104.

Fillmore, C. J. (1968) The Case for Case. In: Bach, E./Harms, R. T. (eds.) (1968) Universals in linguistic Theory. New York: Holt, Rinehart and Winston, Inc, 1-88.

Fillmore, C. J. (1975) An Alternative to Checklist Theories of Meaning. In: Proceedings of the First Annual Meeting of the Berkeley Linguistics Society. Berkeley, 123-131.

Fillmore, C. J. (1976) Frame Semantics and the Nature of Language. In: Harnald, S./Steklis, H./Lancaster, J. (eds.) (1976) Origins and Evolution of Language and Speech. New York: New York Academy of Science, 20-32.

Fillmore, C. J. (1977) The Case for Case Reopened. In: Cole, P./Sadock, J. M. (eds.) (1977) Syntax and Semantics: Grammatical Relations, Bd. V III. New York: Academic Press, 59-82.

Fillmore, C. J. (1977a) Scene-and-Frames Semantics. In: Zampolli, A. (ed.) (1977) Linguistic Structures Processing. Amsterdam/New York: North-Holland, 55-81.

Fillmore, C. J. (1977b) Topic in lexical semantics. In: Cole, R. W. (ed.) (1977) Current Issues in Linguistic Theory. Bloomington et al.: Indiana University Press, 76-138.

Fillmore, C. J. (1982) Frame semantics. In: Linguistics in Morning Calm, 111-137.

Fillmore, C. J. (1984) Lexical Semantics and Text Semantics. In: Copeland, J. E. (ed.) (1984) New Directions in Linguistics and Semiotics. Houston, Texas: Rice University, 123-147.

Fillmore, C. J. (1985) Frames and the Semantics of Understanding. In: Quaderni di Semantika 6, 2, 225-254.

Fischer, K. (2000) From Cognitive Semantics to Lexical Pragmatics. Berlin/New York: de Gruyter.

Fleischer, M., (1989) Die sowjetische Semiotik: Theoretische Grundlage der Moskauer un Tartauer Schule. Tübingen: Stauffenburg.

Fleischer, W. /Bartz, I. (1995) Wortbildung der deutschen Gegenwartssprache. Tübingen: Niemeyer.

Flood, J. (ed.) (1984) Understanding Reading Comprehension: Cognition, Language and the Structure of Prose. Delaware: International Reading Association.

Fodor, J. A. (1983) Modularity of Mind. Cambridge: MIT Press.

Fox, C. (1992) Lexical Analysis and Stoplists. In: Frakes, W., B./ Baeza-Yates, R. (eds.) (1992) Information Retrieval and Data Structures. Upper Saddle River, N. J.: Prentice Hall, 102-130.

Frakes, W., B./ Baeza-Yates, R. (eds.) (1992) Information Retrieval and Data Structures. Upper Saddle River, N. J.: Prentice Hall.

Francis, W. N./Kucera, H. (1979) Brown Corpus Manual. In: http://icame.uib.no/brown/bcm.html (03.04. 2004).

Frants, I. V./Shapiro, J./Voiskunski, V. G. (1997) Automated Information Retrieval: Theory and Methods. San Diego et al.: Academic Press.

Freedle, R. O. (ed.) (1979) New Directions in Discourse Processing. Norwood, N. J.: Publishing Corporation.

Fritz, G. (1982) Kohärenz: Grundfrage der linguistischen Kommunikationsanalyse. Tübingen: Narr.

Gaus, W. (2003) Dokumentations- und Ordnungslehre: Theorie und Praxis des Information Retrieval. Berlin/Heidelberg: Springer.

Garside R./ Leech, G./ Sampson, G. (1987) The Computational Analysis of English: A Corpus-based Approach. London: Longman.

Gašević, D./Djurić, D./Devedžić, V. (2006) Model Driven Architecture and Ontology Development. Berlin/Heildelberg: Springer.

Geckeler, H. (1971) Strukturelle Semantik und Wortfeldtheorie. München: Fink.

Geckeler, H. (ed.) (1978) Strukturelle Bedeutungslehre. Darmstadt: Wissenschaftliche Buchgesellschaft.

Gernsbacher, M. A./Givón, T. (eds.) (1995) Coherence in Spontaneous Text. Amsterdam: Benjamins.

Gernsbacher, M. A./Givón, T. (1995a) Introduction: Coherence as a Mental Entity. In: Gernsbacher, M. A./Givón, T. (eds.) (1995) Coherence in Spontaneous Text. Amsterdam: Benjamins, VII – X.

Givón, T. (1995) Coherence in Text vs. Coherence in Mind. In: Gernsbacher, M. A./Givón, T. (eds.) (1995) Coherence in Spontaneous Text. Amsterdam: Benjamins, 59-115.

Goldstein, E. B. (2002) Wharnehmungspsychologie. Heidelberg: Spektrum Akademischer Verlag.

Gonnet, G./Baeza-Yates, R. (1991) Handbook of Algorithms and Data Structures. Workingham: Addison-Wesley.

Graumann, C. F. (ed.) (1966) Denken. Köln/Berlin: Kiepenheuer/Witsch.

Gregg, L. W./Steinberg, E. R. (eds.) (1980) Cognitive Processes in Writing. Hillsdale, N. J.: Erlbaum.

Green, R. (2001) Relationships in the Organisation of Knowledge: An Overview. In: Bean, C. A./Green, R. (eds.) (2001) Relationships in the Organization of knowledge. Dordrecht et al.: Kluwer Academic Press, 3-18.

Green, R. (2002) Internally–Structured Conceptual Models in Cognitive Semantics. In: Green, R./Bean, C. A./Myaeng, S. H. (eds.) (2002) The Semantics of Relationships: An Interdisciplinary Perspective. Dordrecht et al.: Kluwer Academic Publishers, 73-89.

Green, R./Bean, C. A./Myaeng, S. H. (eds.) (2002) The Semantics of Relationships: An Interdisciplinary Perspective. Dordrecht et al.: Kluwer Academic Publishers.

Grice, H. P. (1975) Logic and conversation. In: Cole, P./Morgan, J. L. (eds.) (1975) Syntax and Semantics: Speech Acts, Bd. III. New York: Academic Press, 41-58.

Grice, H. P. (1981) Presupposition and Conversational Implicature, In: Cole, P. (ed.) (1981) Radical Pragmatics. New York: Academic Press, 183-198.

Grosse, E. U. (1976) Text und Kommunikation. Stuttgart et al.: Kohlhammer.

Grosz, B. J./Jones, K., S./Webber, B. L. (eds.) (1986) Redings in Natural Language Processing. Los Altos: Morgan Kaufmann.

Guinchar, C./Menou, M. (1983) General Introduction to the Techniques of Information and Documentation Work. Paris: Unesco.

Gülich, E./Raible, W. (eds.) (1972) Textsorten: Differenzierungskriterien aus linguistischer sicht. Frankfurt: Athenäum.

Gülich, E./Raible, W. (1977) Linguistische Textmodelle. München: Fink.

Habel, C. (ed.) (1985) Künstliche Intelligenz. Repräsentation von Wissen und natürliche System. Frühjahrschule Dassel. Berlin: Springer.

Habel, C. (1985) Das Lexikon in der Forschung der Künstlichen Intelligenz. In: Schwarze, C./Wunderlich, D. (eds.) (1985) Handbuch der Lexikologie. Königstein: Athnäum, 441-474.

Harberzettl, S./Wegner, H.(eds.) (2003) Spracherwerb und Konzeptualisierung. Frankfurt a. M. et al.: Peter Lang.

Halliday, M. A. K. (1974) The Place of "Functional Sentence Perspective" in the System of Linguistic Description, In: Daneš, F. (ed.) (1974) Papers on Functional Sentence Perspective. The Hague: Mouton, 43-53.

Halliday, M. A. K. /Hasan, R. (1976) Cohesion in English. London/New York: Longman.

Harman, D./Fox, E./Baeza-Yates, R./Lee, W. (1992) Inverted Files. In: Frakes, W., B./ Baeza-Yates, R. (eds.) (1992) Information Retrieval and Data Structures. Upper Saddle River, N. J.: Prentice Hall, 28-43.

Harnald, S./Steklis, H./Lancaster, J. (eds.) (1976) Origins and Evolution of Language and Speech. New York: New York Academy of Science.

Hartung, W. (2000) Kommunikationsorientierte und handlungstheoretisch ausgerichtete Ansätze. In: Brinker, K./Antos, G./Heinemann, W./Sager, S. F. (eds.) (2000) Text- und Gesprächslinguistik, Bd. I. Berlin/New York: de Gruyter, 83-96.

Harweg, R. (1968) Pronomina und Textkonstitution. München: Fink.

Hasan, R. (1984) Coherence and Cohesive Harmony. In: Flood, J. (ed.) (1984) Understanding Reading Comprehension: Cognition, Language and the Structure of Prose. Delaware: International Reading Association, 181-219.

Hayes, J. R./Flower, L. S. (1980) Identifying the Organization of Writing Process. In: Gregg, L. W./Steinberg, E. R. (eds.) (1980) Cognitive Processes in Writing. Hillsdale, N. J.: Erlbaum, 3-30.

Haugeland, L. (ed.) (1981) Mind Design. Cambridge: MIT Press.

Hausser, R. (1998) Drei prinzipielle Methoden der automatischen Wortformerkennung. In: Sprache und Datenverarbeitung 22, 38-57.

Heinemann, W., 2000, Aspekte der Textsortendifferenzierung. In: Brinker, K./Antos, G./Heinemann, W./Sager, S. F. (eds.) (2000) Text- und Gesprächslinguistik, Bd. I. Berlin/New York: de Gruyter, 523-546.

Heit, E. (1997) Knowledge and Concept Learning. In: Lamberts, K./Shanks, D. (eds.) (1997) Knowledgs, Concepts and Categories. East Sussex: Psychology Press, 7-42.

Hellwig, P. (1984) Grundzüge einer Theorie des Textzusammenhangs. In: Rothkegel, A./Sandig, B. (eds.) (1984) Text-Textsorte-Semantik. Hamburg: Buske, 51-79.

Hellwig, P. (1989) Parsing natürlicher Sprache: Realisierung. In: Bátori, I., S./Lenders, W./Putschke, W. (eds.) (1989) Computational Linguistics/Computer-Linguistik. An International Handbook on Computer Oriented Language Research and Application. Berlin/New Y-ork: de Gruyter, 378-432.

Herrmann, T. (2003) Kognitive Grundlage der Sprachproduktion. In: Rickheit, G./Herrmann, T./ Deutsch, W. (eds.) (2003) Psycholinguistik: Ein internationales Handbuch zur Sprach- und Kommunikationswissenshcaft. Berlin/New York: de Gruyter, 228-244.

Hirst, G./St-Onge, D. (1998) Lexical Chains as Representations of Context for the Detection and Correction of Malapropism. In: Fellbaum, C. (ed.) (1998) Wordnet: An Electronic Lexical Database. Cambridge/London: MIT Press, 305-332.

Hoey, M. (1991) Patterns of Lexis in Text. Oxford: Oxford University Press.

Hoffmann, J. (1986) Die Welt der Begriffe: Psychologische Untersuchungen zur Organisation menschlichen Wissens. Weinheim: Psychologie Verlag Union.

Hovy, E./Lin, C.-Y. (1999) Automated Text Summarization in SUMMAIST. In: Mani, I./Maybury, M. T. (eds.) (1999) Advances in Automatic Text Summarization. Cambridge/London: MIT Press, 81-94.

Hörmann, H. (1976) Meinen und Verstehen: Grundzüge einer psychologischen Semantik. Frankfurt a. M.: Suhrkamp.

Hjørland, B. (1997) Information Seeking and Subject Representation: An activity-Theoretical Approach to Information Science. Westport/London: Greenwood Press.

Huddleson, R. (1984) Introduction to the Grammar of English. Cambridge: Cambridge University Press.

Hyldgaard-Jensen, K./Zettersten, A. (eds.) (1985) Symponium on Lexicography: 2. Proceedings of the Second International Symposium on Lexicography May 26-17, 1984 at the University of Copenhagen. Tübingen: Niemeyer.

Isenberg, H. (1970) Der Begriff ‚Text' in der Sprachtheorie. Berlin: Arbeitsgruppe Strukturelle Grammatik (=AGS-Bericht 8).

Isenberg, H. (1977) Text als kommunikative Einheit. In: Viehweger, D. (ed.) (1977) Probleme der semantischen Analyse. Berlin: Akademischer Verlag (= studia grammatica XV), 358-377.

Isenberg, H. (1977a,) ‚Text' versus ‚Satz'. In: Daneš, F./Viehweger, D. (eds.) (1977) Probleme der Textgrammatik 2. Berlin: Akademie Verlag (= studia grammatica XVIII), 119-146.

Isenberg, H. (1984) Texttypen als Interaktionstypen. In: Zeitschrift für Germanistik 5, 261-270.

Jackendoff, R. (1983) Semantics and Cognition. Cambridge: MIT Press.

Jackendoff, R (1990) Semantic Structures. Cambridge: MIT Press.

Jarvella, R. J./Klein, W. (eds) (1982) Speech, Place and Action: Studies in Deixis and Related Topics. Chichester: Willey.

Johnson-Laird, P. N. (1984) Semantic Primitives or Meaning Postulates, Mental Models or Propositional Representations. In: Bara, B. G./Guida, G. (eds.) (1984) Computational Models of Natural Language Processing. Amsterdam/New York: North-Holland, 227-246.

Johnson-Laird, P. N. (1989) Mental Models. In: Posner, M. I. (ed.) (1989) Foundation of Cognitive Science. Cambridge/London: MIT Press, 469-499.

Johnson-Laird, P. N./Wason, P. C. (eds.) (1977) Thinking. Cambridge: Cambridge University Press.

Just, M. A./Carpenter, P. A. (eds.) (1977) Cognitive Processes in Comprehension. Hillsdale: Erlbaum.

Just, M. A./Carpenter, P. A. (1980) A Theory of Reading: From eye Fixations to Comprehension. In: Psychological Review 87, 329-354.

Jünttner, G. Ä/Günzer, U. (1988) Methoden der künstlichen Intelligenz für Information Retrieval. Münschen et al.:Saur.

Kallmeyer, W./Klein, W./Meyer-Hermann, R./Netzer, K./Siebert, H. J. (eds.) (1974) Lektürekollege zur Textlinguistik, Bd. I. Frankfurt: Athenäum.

Kallmeyer, W./Klein, W./Meyer-Hermann, R./Netzer, K./Siebert, H. J. (eds.) (1974a) Lektürekollege zur Textlinguistik, Bd. II. Frankfurt: Athenäum.

Kay, P. (1971) Taxonomy and Semantic Contrast. In: Language 47, 4, 866-887.

Kastovsky, D. (1982) Wortbildung und Semantik, Düsseldorf et al.: Francke.

Kamppinen, M. (ed.) (1993) Consciousness, Cognitive Schemata, and Relativism. Dordrecht: Kluwer Academic Press.

Kamppinen, M. (1993a) Cognitive Schemata. In: Kamppinen, M. (ed.) (1993) Consciousness, Cognitive Schemata, and Relativism. Dordrecht: Kluwer Academic Press, 133-170.

Kebeck, G. (1994) Wahrnehmung: Theorien, Methoden und Forschungsergebnisse der Wahrnehmungspsychologie. Weinheim/München: Juventa.

Kelter, S. (2003) Mentale Modell. In: Rickheit, G./Herrmann, T./ Deutsch, W. (eds.) (2003) Psycholinguistik: Ein internationales Handbuch zur Sprach- und Kommunikationswissenshcaft. Berlin/New York: de Gruyter, 505-517.

Kintsch, W. (1974) The Representation of Meaning in Memory. Hillsdale, N. J.: Erlbaum.

Kintsch, W. (1988) The Role of Knowledge in Discourse Comprehension: A construction- integration model. In: Psychological Review 95,163-182.

Kittay, E. F. /Lehrer, A. (1981) Semantic Fields and the Structure of Metaphor. In: Studies in Language 5, 31-63.

Klause, J. (1996) Principles of Content Analysis for Information Retrieval Systems: An Overview. In: Zuell, C./Harkness, J.,/Hoffmeyer-Zlotnik, J. H. P. (eds.) (1996) ZUMA-Nachrichen Speziel: Textanalysis and Computers. Mannheim: ZUMA, 76-100.

Kleiber, Georges (1993) Prototypen Semantik, Eine Einführung. Tübingen: Narr.

Klein, W./Stutterheim, C. (1987) Quaestio und referentielle Bewegung in Erzählungen. In: Linguistische Berichte 109, 163-183.

Klein, W./Stutterheim, C. (1991) Textstructure and Referential Movement. In: Sprache und Pragmatik 22, 1-31.

Klix, F. (1980) Erwachenes Denken. Berlin: Volk und Wissen.

Klix, F. (1988) Gedächtnis und Wissen. In: Mandl, H./Spada, H. (eds.) (1988) Wissenspsychologie. München/Weinheim: Psychologie Verlag Union, 19-54.

Knorz, G. (1995) Information Retrieval-Anwendungen. In: Kleines Lexikon der Informatik und Wirtschaftsinformatik, 244-248.

Koch, W. A. (1972) Strukturelle Textanalyse. Hildesheim: Olms.

Koch, W. A. / Rosengren, I./Schonebohm, M. (1981) Ein pragmatisch orientiertes Textanalyseprogramm. In: Sprache und Pragmatik (= Lunder Symposium 1980), 155-203.

Konerding, K.-P. (1993) Frame und lexikalisches Bedeutungswissen: Untersuchung zur linguistischen Grundlegung einer Frametheorie und zur ihrer Anwendung in der Lexikographie. Tübingen: Niemeyer.

Kornadt, H.-J./Grabowski, J./Mangold-Allwinn, R. (eds.) (1994) Sprache und Kognition: Perspektiven moderner Sprachpsychologie. Heidelberg et al.: Akademischer Verlag.

Kowalski, G. (1997) Information Retrieval Systems: Theory and Implementation. Boston et al: Kluwer Acamemic Publishers.

Kuhlen, R. (1985) Verarbeitung von Daten, Repräsentation von Wissen, Erarbeitung von Information. Primat der Pragmatik bei informationeller Sprachverarbeitung. In: Endres-Niggemeyer, B./Krause, J., (eds.) (1985) Sprachverarbeitung in Information und Dokumentation, Jahrestagung der GLDV. Berlin et al.: Springer, 1-22.

Lakoff, G. (1987) Women, Fire, and Dangerous Things: What Categories Reveal about the Mind. Chicago/London: University of Chicago Press.

Lamberts, K./Shanks, D. (eds.) (1997) Knowledgs, Concepts and Categories. East Sussex: Psychology Press.

Lancaster, F. W. (1998) Indexing and abstracting: theory and practice. 2. Aufl. London: Library Association Publishing.

Langacker, R. W. (1987) Foundation of Cognitive Grammar, Bd. I. Stanford: Stanford University Press.

Langacker, R. W. (1994) Foundation of Cognitive Grammar, Bd. II. Stanford: Stanford University Press.

Langridge, D. W. (1994) Inhaltsanalyse: Grundlage und Methode. München et al.: Saur.

Lappin, S. (ed.) (1996) The Handbook of Contemporary Semantic Theory. Oxford: Blackwell.

Leech. (1981) Semantics, the Study of Meaning. Harmondsworth: Penguin.

Lehnert, W. G. (1980) The Role of Scripts in Understanding. In: Metzing, D. (ed.) (1980) Frame Conceptions and Text Understanding. Berlin'/New York: de Gruyter, 79-95.

Lehrer, A. (1974) Semantic Fields and lexical Structure. Amsterdam: North-Holland.

Lehrer, A. (2002) Pradigmatic relation of Excusion and Opposition: Gradable and complementarity. In: Cruse, D. A./Hundsnurscher, F./Job, M./Lutzeier, P. R. (eds.) (2002) Lexikologie:

Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wortschätzen, Bd. I. Berlin/New York: de Gruyter, 498-507.

Lehrer, A./Kittay, E. F. (eds.) (1992) Frames, Fields, and Contrast: New Essays in Semantic and Lexical Organisation. Hillsdale, N. J.: Erlbaum.

Le Ny, J. F./Kintsch, W. (eds.) (1982) Language and Comprehension. Amsterdam: North-Holland.

Levelt, W. J. M. (1982) Cognitive Styles in the Use of Spatial Direction Terms. I: Jarvella, R. J./Klein, W. (eds) (1982) Speech, Place and Action: Studies in Deixis and Related Topics. Chichester: Willey, 251-268.

Levelt, W. J. M. (1982a) Linearization in Describing Spatial Network: In: Peters, S./Saarinen, E. (eds.) (1982) Processes, Beliefs and Questions. Dortrecht: Reidel, 199-220.

Levelt, W. J. M. (1989) Speaking: From Intention to Articulation. Cambridge: MIT Press.

Levelt, W. J. M. (1992) Accessing Words in Speech Production: Stages, Process and Representations. In: Cognition 42, 1-22.

Levi, J. N. (1978) The Syntax and Semantics of Complex Nominals. New York: Academic Press.

Levin, B./Pinker, S. (eds.) (1992) Lexical & Conceptual Semantics. Cambridge: Blackwell.

Levinson, S. C. (1997) From Outer to Inner Space: Linguistic Categores and Nonlinguistic Thinking´. In: Nuyts, J./Perterson, E. (eds.) (1997) Language and Conceptualization. New York: Cambridge University Press, 13-45.

Lewis, D. (1975) Konventionen: Eine sprachphilosophische Abhandlung. Berlin/New York: de Gruyter.

Lin, C. Y. (1995) Knowledge-based Automatic Topic Identification. In: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics(ACL-95), 308-320.

Linsky, L. (ed.) (1987) Semantics and the philosophy of language. Urbana: University of Illinois Press.

Lipka, L. (1987) Protptype semantics or feature semantics: An alternative? Lörscher, W./Schulze, R. (eds.) (1987) Perspectives on Language in Performance. Tübingen: Narr, 282-298.

Lipka, L. (2002) English Lexicology: Lexical Structure, Word Semantics & Word-formation. Tübingen: Narr.

Litowitz, B. E. /Novy, F. A. (1984) Expression of the Part-Whole Semantic Relation by 3- to 12-years old Children. In: Journal of Child Language 11, 159-178.

Lutzeier, P. R. (1981) Wort und Feld. Tübingen: Niemeyer.

Lutzeier, P. R. (1985) Linguistische Semantik. Stuttgart: Metzler.

Löbner, S. (2003) Semantik: Eine Einführung. Berlin/New York: de Gruyter.

Lörscher, W./Schulze, R. (eds.) (1987) Perspectives on Language in Performance. Tübingen: Narr.

Lötscher, A. (1983) Satzakzent und funktionale Satzperspektive im Deutschen. Tübingen: Niemeyer.

Lötscher, A. (1984) Satzgliederung und funktionale Satzperspektive. In: Stickel, G. (ed.) (1984) Pragmatik in der Grammatik. Düsseldorf: Pädagogischer Verlag (=Jahrbuch 1983 des Instituts für deutsche Sprache), 118-151.

Lundquist, L./Jarvella R. J. (eds) (2000) Language, Text and Knowledge: Mental Models of Expert Communication. Berlin/New York: de Gruyter.

Lüdi, G. (1985) Zur Zerlegbar von Wortbedeutung. In: Schwarze, C./Wunderlich, D., (eds.) (1985) Handbuch der Lexikologie. Königstein: Athnäum, 64-102.

Luger, G. F. (2002) Künstliche Intelligenz: Strategie zur Lösung komplexer Probleme. München: Pearson.

Lyons, J. (1968) Einführung in die moderne Linguistik. München: Beck.

Lyons, J. (1977) Semantik, Bd. I. München: Beck.

Lyons, J. (1977a) Semantik, Bd. II. München: Beck.

Lyons, J. (1995) Linguistic semantics: An Introduction. Cambridge: Cambridge University Press.

Lyons, J. (2002) Sense Relations: An overview. In: Cruse, D. A./Hundsnurscher, F./Job, M./Lutzeier, P. R. (eds.) (2002) Lexikologie: Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wortschätzen, Bd. I. Berlin/New York: de Gruyter, 466-472.

Mandl, H. (ed.) (1981) Zur Psychologie der Textverarbeitung: Ansätze, Befunde, Probleme. München: Urban & Schwanzenberg.

Mandl, H./Friedrich, H. F./Hron, A. (1988) Theoretische Ansätze zum Wissenserwerb. In: Mandl, H./Spada, H. (eds.) (1988) Wissenspsychologie. München/Weinheim: Psychologie Verlag Union, 123-160.

Mandl, H./Spada, H. (eds.) (1988) Wissenspsychologie. München/Weinheim: Psychologie Verlag Union.

Mani, I. (2001) Automatic Summarization. Amsterdam/Philadelphia: Benjamins.

Mani, I./Maybury, M. T. (eds.) (1999) Advances in Automatic Text Summarization. Cambridge/London: MIT Press.

Manning, C. D./Schütze, H. (2000) Foundation of Statistical Natural Language Processing. Cambridge: MIT Press.

Marcu, D. (1997) From Discourse Structure to Text Summaries. In: Proceedings of the Workshop on Intelligent Scalable Text Summarization. Madrid: Association for Computational Linguistics, 82-88.

Markman, E. M. (1989) Categorization and Naming in Children. Cambridge: MIT Press.

Markman, E. M. (1990) Constraints Children Place on Word Meanings. In: Cognitive Science 14, 57-77.

Markowitsch, H. J. (1996) Neuropsychologie des menschlichen Gedächtnisses. In: Spektrum der Wissenschaft 09, 52-61.

Martinich A. P. (ed.) (1985) The Philosophy of Language. New York/Oxford: Oxford University Press.

Mathesius, V. (1929) Zur Satzperspektive im modernen Englisch. In: Archiv für das Dtudium der neueren Sprachen und Literaturen 84, 202-204.

Meadow, C. T./Boyce, B./Kraft, D. (2000) Text Information Retrieval Systems. San Diego et al.: Academic Press.

Meibauer, J. (1999) Pragmatik: Eine Einführung. Tübingen: Stauffenburg.

Meibauer, J/Rothweiler, M. (eds.) (1999) Das Lexikon im Spracherwerb. Tübingen/Basel: Francke.

Mervis, C. B./Pani, J. R. (1980) Acquisition of Basic Objects. In Cognitive Psychology 12, 496-522.

Metzger, W. (1954) Grundbegriff der Gestaltpsychologie. In Metzger, W. (1986) Gestalt-Psychologie: Ausgewählte Werke aus den Jahren 1950 bis 1982. Frankfurt a. M.: Kramer, 124-133.

Metzger, W. (1968) Psychologie. Darmstadt: Steinkoff.

Metzger, W. (1986) Gestalt-Psychologie: Ausgewählte Werke aus den Jahren 1950 bis 1982. Frankfurt a. M.: Kramer.

Metzing, D. (ed.) (1980) Frame Conceptions and Text Understanding. Berlin/New York: de Gruyter.

Metzing, D. (1981) Frame Representation and Lexical Semantics, In: Eikmeyer, H.-J./Riesler, H. (eds.) (1981) Words, Words and Contexts: New Approaches in Word Semantics. Berlin/New York: de Gruyter, 320-342.

Michaels, C. F. /Carello, C. (1981) Direct Perception. Englewood: Prentice Hall.

Miller, G. A. (1998) Nouns in WordNet. In: Fellbaum, C. (ed.) (1998) Wordnet: An Electronic Lexical Database. Cambridge/London: MIT Press, 23-46.

Miller, G. A./Johnsohn-Laird, P. N. (1976) Perception and Language. Cambridge: Cambridge University Press.

Miller, G. A./Beckwith, R./Fellbaum, C. D./Gross, D./Miller, K. (1993) Introduction to WordNet. In: http://wordnet.princeton.edu/5papers.pdf (01.06.2004).

Miller, G. A./Fellbaum, C. (1992) Semantic Networks of English. In: Levin, B./Pinker, S. (eds.) (1992) Lexical & Conceptual Semantics. Cambridge: Blackwell, 197-230.

Minsky, M. (1975) A Famework for Representing Knowledge. In Winston, P. H. (1975) The Psychology of Computer Vision. New York: McGraw-Hill, 211-278.

Minsky, M. (1977) Frame-System Theory. In: Johnson-Laird, P. N./Wason, P. C. (eds.) (1977) Thinking. Cambridge: Cambridge University Press, 355-376.

Minsky, M. (1980) Famework for Representing Knowledge.(leicht gekürzte und veränderte Fassung von Minsky 1975). In: Metzing, D. (ed.) (1980) Frame Conceptions and Text Understanding. Berlin/New York: de Gruyter, 1-25.

Minsky, M. (1981) Famework for Representing Knowledge.(leicht veränderte Fassung von Minsky 1975). In: Haugeland, L. (ed.) (1981) Mind Design. Cambridge: MIT Press, 95-128.

Mitkov, R. (2002) Anaphora Resolution. London et al.: Longman.

Moens, M.-F. (2000) Automatic Indexing and Abstracting of Document Texts. Boston et al.: Kluwer Academy Publishers.

Moltmann, F. (1997) Part and Whole in Semantics. Oxford: Oxford University Press.

Montada, L. (2002) Die geistige Entwicklung aus der Sicht Jean Piagets. In: Oerter, R./Montada, L. (eds.) (2002) Entwicklungspsychologie. Weinheim et al.: Beltz, 418-435.

Morris, J./Hirst, G. (1991) Lexical Cohesion Computed Thesaural Relations as an Indicator of the Structure of Text. In: Computational Linguistics 17, 1, 21-48.

Motsch, W. (1987) Zur Illokutionsstruktur von Feststellungssatz. In: Phonetik, Sprachwissenschaft und Kommunikationsforschung 40, 45-67.

Motsch, W. (2000) Handlungsstrukturen von Texten. In: Brinker, K./Antos, G./Heinemann, W./Sager, S. F. (eds.) (2000) Text- und Gesprächslinguistik, Bd. I. Berlin/New York: de Gruyter, 414-422.

Motsch, W./Pasch, R. (eds.) (1987) Satz, Text, sprachliche Handlung. Berlin: Akademie Verlag (=Studia grammatica XXV).

Motsch, W./Pasch, R. (1987a) Illokutive Handlung. In: Motsch, W./Pasch, R. (eds.) (1987) Satz, Text, sprachliche Handlung. Berlin: Akademie Verlag (=Studia grammatica XXV), 11-79.

Motsch, W./ Viehweger, D. (1981) Sprachhandlung, Satz und Text. In: Sprach und Pragmatik: Lundner Syposium 1980, 125-154.

Motsch, W./ Viehweger, D. (eds.) (1983) Richtung der modern Semantikforschung. Berlin: Akademie Verlag.

Munitz, M. K., Unger, P. K. (eds.) (1974) Semantics and Philosophy. New York: New York University Press.

Munske, H. H./Polenz, P./Reichmann, O./Hildebrandt, R. (eds.) (1988) Deutscher Wortschatz. Berlin: de Gruyter.

Murphy, G./Lasaline, M. (1997) Hierarchical Structure in Concepts and the Basic Level of Categorization. In: Lamberts, K./Shanks, D. (eds.) (1997) Knowledgs, Concepts and Categories. East Sussex: Psychology Press, 93-132.

Neisser, U. (1976) Cognition and Reality: Principles and Implication of Cognitive Psychology. San Francisco: Freeman.

Nelson, K. (1983) The Conceptual Base for Language. In: Seiler, T. B./Wannenmacher, W. (eds.) (1983) Concept Development and the Development of Word Meaning. Berlin/Heidelberg: Springer, 173-188.

Nohr, H. (2001) Automatische Indexierung: Einführung in betriebliche Verfahren, System und Anwendungen. Potsdam: Verlag für Berlin-Brandenburg.

Nussbaumer, M. (1991) Was Texte sind und Wie sie sein sollen: Ansätze zu einer sprachwissenschaftlichen Begründung eines Kriterienrasters zur Beurteilung von schriftlichen Schülertexten. Tübingen: Niemeyer.

Nuyts, J./Perterson, E. (eds.) (1997) Language and Conceptualization. New York: Cambridge University Press.

Oakhill, J./Garnham, A. (eds.) (1996) Mental Models in Cognitive Science. East Sussex: Psychology Press.

Oates, J./Grayson, A. (eds.) (2004) Cognitive and Language Development in Children. Oxford: Blackwell.

Oerter, R./Montada, L. (eds.) (2002) Entwicklungspsychologie. Weinheim et al.: Beltz.

Palmer, S. E. (1978) Fundamental Aspects of Cognitive Representation. In: Rosch, E./Lloyd, B. B. (eds.) (1978) Cognition and Categorization. Hillsdale, N. J.: Erlbaum, 259-303.

Panther, K.-U./Radden, G. (eds.) (1999) Metonymy in Language and Thought. Amsterdam/Philadelphia: Benjamins.

Peters, S./Saarinen, E. (eds.) (1982) Processes, Beliefs and Questions. Dortrecht: Reidel.

Piaget, J. (1983) Dialogues on the Psychology of Thought. In: Rieber, R. W. (1983) Dialogues on the Psychology of Language and Thought. New York: Plenum Press, 107-125.

Pinkal, M. (1985) Kontextabhängigkeit, Vagheit, Mehrdeutigkeit. In: Schwarze, C./Wunderlich, D. (eds.) (1985) Handbuch der Lexikologie. Königstein: Athenäum, 27-59.

Pinto Molina, M. (1995) Document abstracting: Toward a methodological model. In: Journal of the American Society for Information Science 46(3), 225-234.

v. Polenz, P. (1988) Deutsche Satzsemantik, Grundbegriffe des Zwischen-den-Zeilen-Lesens. Berlin/New York: de Gruyter.

Posner, M. I. (1986) Empirical Studies of Prototypes. In: Craig, C. (ed.) (1986) Noun Classes and Categorization. Amsterdam: Benjamins, 53-61.

Posner, M. I. (ed.) (1989) Foundation of Cognitive Science. Cambridge/London: MIT Press.

Postman, L./Keppel, G. (eds.) (1970) Norms of Word Association. New York: Acamedic Press.

Pribbenow, P (2002) Meronymic Relationships. In: Green, R./Bean, C. A./Myaeng, S. H. (eds.) (2002) The Semantics of Relationships: An Interdisciplinary Perspective. Dordrecht et al.: Kluwer Academic Publishers, 35-50.

Pulman, S. G. (1983) Word Meaning and Belief. London: Croom Helm.

Putnam, H. (1975) Mind, Language and Reality: Philosophical Papers, Bd. II. Cambridge: Cambridge University Press.

Putnam, H. (1979) Die Bedeutung von Bedeutung. Frankfurt: Klostermann.

Quasthoff, U. M. (1985) Textverstehen und Textproduktion. In: Habel, C. (ed.) (1985) Künstliche Intelligenz. Repräsentation von Wissen und natürliche System. Frühjahrschule Dassel. Berlin: Springer, 184-248.

Quinn, P. C./Oates, J. (2004) Early Category Representation and Concepts. In: Oates, J./Grayson, A. (2004) Cognitive and Language Development in Children. Oxford: Blackwell, 21-60.

Quirk, R./Greenbaum, S./Leech, J./Svartvik, J. (1972) A Grammar of Contemporary English. New York/London: Seminar Press.

Quirk, R./Greenbaum, S./Leech, J./Svartvik, J. (1985) A comprehensive grammar of English language. New York/London: Longman.

Radden, G./Kövecses, Z. (1999) Toward a Theory of Metonymy. In: Panther, K.-U./Radden, G. (eds.) (1999) Metonymy in Language and Thought. Amsterdam/Philadelphia: Benjamins, 17-59.

Raible, W. (1981) Von der Allgegenwart des Gegensinns. In: Zeitschrift für romanische Philologie 97, 1-40.

Reimer, U. (1991) Einführung in die Wissensrepräsentation. Stuttgart: Teubner.

Rickheit, G. (ed.) (1991) Kohärenzprozesse. Opladen: Westdeutscher Verlag.

Rickheit, G. (2000) Kohärenz und Kohäsion. In: Brinker, K./Antos, G./Heinemann, W./Sager, S. F. (eds.) (2000) Text- und Gesprächslinguistik, Bd. I. Berlin/New York: de Gruyter, 275-283.

Rickheit, G., /Bock, M. (eds.) (1983) Psycholinguistic Studies in Language Processing. Berlin/New York: de Gruyter.

Rickheit, G. /Habel, C. (eds.) (1995) Focus and Coherence in Discourse Processing. Berlin/New York: de Gruyter.

Rickheit, G./Schade, U. (2000) Kohärenz und Kohäsion. In: Brinker, K./Antos, G./Heinemann, W./Sager, S. F. (eds.) (2000) Text- und Gesprächslinguistik, Bd. I. Berlin/New York: de Gruyter, 275-283.

Rickheit, G./Herrmann, T./ Deutsch, W. (eds.) (2003) Psycholinguistik: Ein internationales Handbuch zur Sprach- und Kommunikationswissenshcaft. Berlin/New York: de Gruyter.

Rickheit, G./Strohner, H. (eds.) (1985) Inferences in Text Processing. Amsterdam: North-Holland.

Rickheit, G./ Strohner, H. (1993) Grundlage der kognitiven Sprachverarbeitung: Modelle, Methode, Ergebnisse. Tübingen/Basel: Francke.

Rickheit, G./ Strohner, H. (1992) Toward a Cognitive Theory of Linguistic coherence. In: Theoretical Linguistics, 209-237.

Rickheit, G./ Strohner, H./Schnotz, W./Strohner H. (1985) The Concept of inference in Discourse Comprehension. In: Rickheit, G./Strohner, H. (eds.) (1985) Inferences in Text Processing. Amsterdam: North-Holland, 3-49.

Rieber, R. W. (1983) Dialogues on the Psychology of Language and Thought. New York: Plenum Press.

Riesler, H. (1977) On the development of Text Grammar, In: Dressler, W. (ed.) (1977) Current Trends in Textlinguistics. Berlin/New York: de Gruyter, 6-20.

Rips, L./Estin, P. (1998) Component of Objects and Evens. In: Journal of Memory and Language 39, 309-330.

Robertson, S. E./ Sparck Jones, K. (1976) Relevance weighting of search terms. In: Journal of the American Society for Information Science 27, 3, 129-146.

Robins, R. H. (1971) General Linguistics. London: Longman.

Roche, E./ Schabes, Y. (1997) Finite-State Language Processing. Cambridge: MIT Press.

Rolf, E. (1982) Sprachliche Informationshandlung. Göppingen: Kümmerle.

Rolf, E. (1993) Die Funktion der Gebrauchstextsorten. Berlin/ New York: de Gruyter.

Rolf, E. (2000) Textuelle Grundfunktion. In: Brinker, K./Antos, G./Heinemann, W./Sager, S. F. (eds.) (2000) Text- und Gesprächslinguistik, Bd. I. Berlin/New York: de Gruyter, 422-435.

Rosch, E. (1975) Cognitive representation of semantic categories. In: Journal of Experimental Psychology 104, 192-233.

Rosch, E. (1975a) The Nature of Mental Code for Color Categories. In: Journal of Experimental Psychology 104, 303-322.

Rosch, E. (1977) Human Categorization. In: Warren, N. (ed.) (1978) Studies in Cross-cultural Psychology, Bd. I. London et al.: Academic Press, 1-49.

Rosch, E. (1978) Principles of categorization, In: Rosch, E./Lloyd, B. B. (eds.) (1978) Cognition and Categorization. Hilsdale, N. J.: Erlbaum, 27-48.

Rosch, E./Lloyd, B. B. (eds.) (1978) Cognition and Categorization. Hilsdale, N. J.: Erlbaum.

Rosch, E./ Marvis, C. B. (1975) Family Resemblance. In: Cognitive Psychology 7, 573-605.

Rosenberg, S. T. (1980) Frame-based Text Processing. In: Metzing, D. (ed.) (1980) Frame Conceptions and Text Understanding. Berlin/New York: de Gruyter, 96-119.

Roth, G. (1996) Das Gehirn und seine Wirklichkeit. Frankfurt a. M.: Suhrkamp.

Rothkegel, A./Sandig, B. (eds.) (1984) Text-Textsorte-Semantik. Hamburg: Buske.

Rothweiler, M (2003) Die Taxonomieannahme im lexikalischen Erwerb. Ergebnisse aus einer empirischen Studie mit sprachnormalen und sprachauffälligen Kindern. In: Harberzettl, S./Wegner, H.(eds.) (2003) Spracherwerb und Konzeptualisierung. Frankfurt a. M. et al. : Peter Lang, 49-70.

Rothweiler, M./ Meibauer, J. (1999) Das Lexikon im Spracherwerb: Ein Überblick. In: Meibauer, J/Rothweiler, M. (eds.) (1999) Das Lexikon im Spracherwerb. Tübingen/Basel: Francke, 9-31.

Rumelhart, D. E. (1975) Notes on a schema for stories. In: Bobrow, A. M./Collins, A. M. (eds.) (1975) Representation and Understanding. New York: Academic Press, 211-236.

Rumelhart, D. E. (1977) Introduction to Human Information Processing. New York: John Wiley & Sons.

Rumelhart, D. E. (1980) Schemata: The Building Blocks of Cognition. In: Spiro, R. J./Bruce, B. C./Brewer, W. F. (eds.) (1980) Theoretical Issues in Reading Comprehension. Hillsdale, N. J.: Erlbaum, 33-58.

Rumelhart, D. E./Linsay, P. H./Norman, D. A. (1972) A Process Model for Long Term Memory. In: Tulving, E./Donalson, W. (eds.) (1972) Organisation of Memory. New York/London: Academic Press, 198-248.

Rumelhart, D. E /Norman, D. A. (1978) Accrection, Tunning and Restructing: Three modes of learning. In: Cotton, J. W./Klatzky, R. L. (eds.) (1978) Semantic Factors in Cognition. Hillsdale, N. J.: Erlbaum, 37-53.

Russel, S./Norvig, P. (2003) Artificial Intelligence: A modern Approach. Upper Saddle River, N. J. : Prentice Hall (dt. Übers. Künstliche Intelligenz: Ein moderner Ansatz. Müchen: Pearson , 2004).

Ružička, R./Motsch, W. (eds.) (1983) Untersuchungen zur Semantik. Berlin: Akademie Verlag (= studia grammatica XXII).

Salton, G. (1989) Automatic Text Processing: The Transformation, Ananylsis and Retrieval of Information by Computer. Readind, MA: Addison-Wesley.

Salton, G./Buckley, C. (1988) Term-Weighting Approaches in Automatic Text Retrieval. In: Information Processing and Management 24, 513-523.

Salton, G/Singhal, A./Mitra, M./Buckely, C. (1999) Automatic Text Structuring and Summarization. In: Mani, I./Maybury, M. T. (eds.) (1999) Advances in Automatic Text Summarization. Cambridge/London: MIT Press, 341-355.

Sandig, B. (1972) Zur Differenzierung gebrauchssprachlicher Textsorten im Deutschen. In: Gülich, E./Raible, W. (eds.) (1972) Textsorten: Differenzierungskriterien aus linguistischer sicht. Frankfurt: Athenäum, 113-124.

Schade, U./Langer, H./ Rutz, H./ Sichelschmidt, L. (1991) Kohärenz als Prozess. In: Rickheit, G. (ed.) (1991) Kohärenzprozesse. Opladen: Westdeutscher Verlag, 7-58.

Schacter, D. L. (1989) Memory. In: Posner, M. I. (ed.) (1989) Foundation of Cognitive Science. Cambridge/London: MIT Press, 683-721.

Schank, R. C. (1975) Conceptual Information Processing. Amsterdam: North-Holland.

Schank, R. C./Abelson, R. P. (1977) Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures. Hillsdale N. J.: Erlbaum.

Scherner, M. (1984) Sprache als Text: Ansätze zu einer sprachwissenschaftliche begründeten Theorie des Textverstehen. Tübingen: Niemeyer.

Scherner, M. (1989) Zur kognitionswissenschaftlichen Modellierung des Textverstehens. In: Zeitschrift für germanistische Linguistik 17, 94-102.

Scherner, M. (2000) Kognitionswissenschaftliche Methoden der Textanalyse. In: Brinker, K./Antos, G./Heinemann, W./Sager, S. F. (eds.) (2000) Text- und Gesprächslinguistik, Bd. I. Berlin/New York: de Gruyter, 186-195.

Schindler, W. (2002) Lexik, Lexikon, Wortschatz. I Cruse, D. A./Hundsnurscher, F./Job, M./Lutzeier, P. R. (eds.) (2002) Lexikologie: Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wortschätzen, Bd. I. Berlin/New York: de Gruyter, 34-44.

Schippan, T. (2002) Lexikologie der deutschen Gegenwartsprache. Tübingen: Niemeyer.

Schmidt, S. J. (1991) Grundriss der empirischen Literaturwissenschaften. Franfurt a. M.: Suhrkamp.

Schnotz, W. (1988) Textverstehen als Aufbau mentaler Modelle. In: Mandl, H./Spada, H. (eds.) (1988) Wissenspsychologie. München/Weinheim: Psychologie Verlag Union, 299-330.

Schnotz, W. (1994) Aufbau von Wissensstruktur: Untersuchungen zur Kohärenzbildung bei Wissenserwerb mit Texten. Weinheim: Psychologie Verlags Union.

Schnotz, W./Ballstaedt, S.-P./Mandl, H. (1981) Kognitive Prozesse beim Zusammenfassen Lehrtexten. In: Mandl, H. (ed.) (1981) Zur Psychologie der Textverarbeitung: Ansätze, Befunde, Probleme. München: Urban & Schwanzenberg, 201-225.

Schwarz, M. (1992) Einführung in die Kognitive Linguistik. Tübingen: Francke.

Schwarz, M. (ed.) (1994) Kognitive Semantik. Tübingen: Narr.

Schwarz, M. (2002) Konzeptuelle Ansätze II: Einebene – Ansatz vs. Mehrebenen - Ansatz. In: Cruse, D. A./Hundsnurscher, F./Job, M./Lutzeier, P. R. (eds.) (2002) Lexikologie: Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wortschätzen, Bd. I. Berlin/New York: de Gruyter, 277-284.

Schwarz, M./Chur, J. (1993) Semantik: Ein Arbeitsbuch. Tübingen: Narr.

Schwarze, C. (1982) Stereotyp und lexikalische Bedeutung. In: Studium Linguistik 13, 1-16.

Schwarze, C. (1988) Textverstehen und lexikalisches Wissen. In: Stechow, A./Schepping, M.-T. (eds.) (1988) Fortschritte in der Semantik, Sonderforschungsberichte. Weinheim: VCA, Acta Humaniora, 139-157.

Schwarze, C./Wunderlich, D. (eds.) (1985) Handbuch der Lexikologie. Königstein: Athnäum.

Schwatz, S. P. (1979) Natural Kind Terms. In: Cognition 7, 301-315.

Schwatz, S. P. (1980) Natural Kinds and Nominal Kinds. In: Mind 89, 182-195.

Searle, J. R. (1969) Speech Acts: An Essay in the Philosophy of Language. Cambridge: Cambridge University Press (dt. Übers. Sprechakte: Ein sprachphilosophischer Essay. Frankfurt a. M.: Suhrkamp, 1971).

Seel, N. M. (2000) Psychologie des Lernens. München/Basel: UTB.

Seiler, T. B. (1984) Begriffsentwicklung und die Veränderung des Verstehens. In: Engelkamp, J. (ed.) (1984) Psychologische Aspekte des Verstehens. Berlin/Heidelberg: Springer (= Lehr und Forschungstexte Psychologie 10), 55-74.

Seiler, T. B. (1985) Sind Begriffe Aggregate von Komponenten oder idiosynkratische Minitheorien?: Kritische Überlegungen zum Komponentenmodell von Dedre Gentner und Vorschläge zu einer alternativen Konzeption. In: Wannenmacher/Seiler 1985, 105-131.

Seiler, T. B./Wannenmacher, W. (eds.) (1983) Concept Development and the Development of Word Meaning. Berlin/Heidelberg: Springer.

Seiler, T. B /Wannenmacher, W. (1983a) How Can We Assess Meaning and Investigate Meaning Development: Theoretical and Methodological Considerations from an Epistemological View. In: Seiler, Th. B. /Wannenmacher, W. (eds.) (1983) Concept Development and the Development of Word Meaning. Berlin/Heidelberg: Springer, 320-339.

Seiler, T. B /Wannenmacher, W. (eds.) (1985) Begriffs- und Wortbedeutungsentwicklung: Theoretische, methodische und methodische Untersuchung. Berlin/Heidelberg: Springer.

Shastri, L. (1988) Semantic Networks: An Evidential Formulation And Its Connectionist Realization. Santa Monteo: Morgan Kaufmann.

Shopen, T. (ed.) (1985) Language Typology and Syntactic Description, Bd. III: Grammatical Categories and the Lexicon. Cambridge: Cambridge University Press.

Sinclair, J. (1966) Beginning the Study of Lexis, In: Bazell, C. E./Catford, J. C./Halliday, M. A. K./ Robins, R. H. (eds.) (1966) In Memory of J. R. Firth. London: Longman, 410-430.

Sörgel, D. (ed.) (1985) Organizing Information. Orlando et al.: Academic Press.

Sowa, J. F. (1984) Conceptual Structures: Information Processing in Mind and Machine. MA: Addison-Wesley.

Sparck Jones, K. (1986) Synonymy and Semantic Relations. Edinburgh: Edinburgh Press.

Sparck Jones, K. (1999) Automatic Summarizing: Factor and Directions. In: Mani, I./Mabury, M. T. (eds.) (1999) Advances in Automatic Text Summarization, Cambridge London: The MIT Press, 1-12.

Spark Jones, K./Kay, M. (1976) Linguistik und Informationswissenschaft. München: UTB.

Sparck Jones, K./Galliers, J. R. (1996) Evaluating Natural Language Processing: An Analysis and Review. New York: Springer.

Spinner, H. F. (1994) Die Wissensordnung: Ein Leitkonzept für die dritte Grundordnung des Informationszeitalters. Opladen: Leske/Budrich.

Spiro, R. J./Bruce, B. C./Brewer, W. F. (eds.) (1980) Theoretical Issues in Reading Comprehension. Hillsdale, N. J.: Erlbaum.

Stairmand, M. A. (1997) Textual Context Analysis for Information Retrieval. In: Proceedings of the 20th Annual ACM SIGIR Conference on Reaserch and Development in IR, 140-147.

Stairmand, M. A./Black, W. J. (1997) Conceptual and contextual Indexing using WordNet-derived lexical chains. In: Proceedings of BCS IRSG Colloquium on Information Retrieval, 47-65.

Stechow, A./Schepping, M.-T. (eds.) (1988) Fortschritte in der Semantik, Sonderforschungsberichte. Weinheim: VCA, Acta Humaniora.

Staab, S./Studer, R. (eds.) (2004) Handbook of Ontologes. Berlin/Heidelberg: Springer.

Stanford, A. J. /Carrod, S. C. (1982) Toward a Psychological Model of Writen Discourse Comprehensio. In: Le Ny, J. F./Kintsch, W. (eds.) (1982) Language and Comprehension. Amsterdam: North-Holland, 147-156.

Steinberg, D. D./Jakobovits, L. A. (1971) Semantics: An interdisciplinary reader in philosophy, linguistics and psychology. Cambridge: Cambridge University Press.

Steiner, E. (2000) Der britische Kontextualismus. In: Brinker, K./Antos, G./Heinemann, W./Sager, S. F. (eds.) (2000) Text- und Gesprächslinguistik, Bd. I. Berlin/New York: de Gruyter, 60-64.

Steinitz, R. (1974), Nominale Pro-Formen. In: Kallmeyer, W./Klein, W./Meyer-Hermann, R./Netzer, K./Siebert, H. J. (eds.) (1974a) Lektürekollege zur Textlinguistik, Bd. II. Frankfurt: Athenäum, 246-265.

Stevenson, R. J. (1996) Mental Models, Propositions, and Comprehension of Pronouns. In: Oakhill, J./Garnham, A. (eds.) (1996) Mental Models in Cognitive Science. East Sussex: Psychology Press, 53-76.

Stickel, G. (ed.) (1984) Pragmatik in der Grammatik. Düsseldorf: Pädagogischer Verlag (=Jahrbuch 1983 des Instituts für deutsche Sprache).

Stock, W. (2007) Information Retrieval: Information suchen und finden. München et al.: Oldenbourg.

Stokes, N. (2004) Application of Lexical Cohesion and Analysis in the Topic Detection and Tracking Domain. Dissertation, University College Dublin.

Strohner, H. (1990) Textverstehen: Kognitive und kommunikative Grundlage der Sprachverarbeitung. Opladen: Werstdeutscher Verlag.

Strohner, H. (1995) Kognitive Systeme: Eine Einführung in die Kognitionswissenschaft. Opladen: Westdeutscher Verlag.

Strohner, H. (2003) Kognitive Voraussetzungen: Wissenssystem-Wissensstruktur-Gedächtniss. In: Rickheit, G./Herrmann, T./ Deutsch, W. (eds.) (2003) Psycholinguistik: Ein internationales Handbuch zur Sprach- und Kommunikationswissenshcaft. Berlin/New York: de Gruyter, 261-274.

Strohner, H. (2003a) Parsing-Prozesse. In: Rickheit, G./Herrmann, T./ Deutsch, W. (eds.) (2003) Psycholinguistik: Ein internationales Handbuch zur Sprach- und Kommunikationswissenshcaft. Berlin/New York: de Gruyter, 524-532.

Sucharowski, W. (1996) Sprache und Kognition. Opladen: Westdeutscher Verlag.

Tannen, D. (1979) What's on a Frame?: Surface Evidence for Underlying Expectation. In: Freedle, R. O. (ed.) (1979) New Directions in Discourse Processing. Norwood, N. J.: Publishing Corporation, 137-182.

Talmy, L. (1985) Lexicalization Patterns: Semantic Structure in Lexicon fForms. In: Shopen, T. (ed.) (1985) Language Typology and Syntactic Description, Bd. III: Grammatical Categories and the Lexicon. Cambridge: Cambridge University Press, 57-149.

Tattersall, I. (2002) Wie der Menschen Denken lernte. In: Spektrum der Wissenschaft, 2002, 4, 56-63.

Tergan, S.-O. (1989) Psychologische Grundlagen der Erfassung individueller Wissensrepräsentationen, Teil 1: Methodologische Aspekte. In: Sprach und Kognition 8, 4, 193-202.

Tergan, S.-O. (1989a) Psychologische Grundlagen der Erfassung individueller Wissensrepräsentationen, Teil 2: Methodologische Aspekte. In: Sprach und Kognition 8, 4, 193-202.

Thaller, M. (1989) Query Net I/O. Göttingen: Max Planken Institut (=Halbgraue Reihe zur historischen Fachinformatik. Serie B: Softwarebeschreibung, Bd. II).

Thorndyke, P. (1977) Cognitive Structures in Comprehension and memory of Narrative Discourse. In: Cognitive Psychology 9, 77-110.

Treisman, A. M./Gelade, G. (1980) A Feature-Integration Theory of Attention. In: Cognitive Psychology 12, 97-136.

Tsohatzidis, S., L. (ed.) (1990) Meaning and Prototypes: Studies in Linguistic Categorization. London: Routledge.

Turner, A./Green, E. (1977) Construction and Use of a Propositional Text Base., University of Colorado: Boulder (=Technical report 63. Institute for the Study of Intellectual Behavior).

Tulving, E. (1972) Episodic and Semantic Memory. In: Tulving, E./Donalson, W. (eds.) (1972) Organisation of Memory. New York/London: Academic Press, 382-403.

Tulving, E./Donalson, W. (eds.) (1972) Organisation of Memory. New York/London: Academic Press.

Tversky, B. (1989) Parts, Partomymies, and Taxonymies. In: Psychological Review 84, 327-352.

Tversky, B. (1990) Where Partonomies and Taxonomies meet. In: Tsohatzidis, S., L. (ed.) (1990) Meaning and Prototypes.: Studies in Linguistic Categorization. London: Routledge, 334-344.

Vater, H. (1994) Einführung in die Textlinguistik. München: Fink.

Vickery, B./Vickery, A. (1989) Information science in Theory and Practice. London: Bowker-Saur.

Viehweger, D. (ed.) (1977) Probleme der semantischen Analyse. Berlin: Akademie Verlag (= studia grammatica XV).

Viehweger, D. (1991) Die Vielfalt textlinguistischer Forschungsansätze – methodologisches Dilemma oder notwendiger Pluralismus? In: Linguistische Studien 209, 200-211.

Voorhees, E. M. (1998) Using WordNet for Text Retrieval. In: Fellbaum, C. (ed.) (1998) Wordnet: An Electronic Lexical Database. Cambridge/London: MIT Press, 285-304.

Wannenmacher, W./Seiler, T. B. (1983) How Can We Assess Meaning and Investigate Meaning Development: theoretical and Methodological Consideration from an Epistemological Point of View. In: Seiler, T. B./Wannenmacher, W. (eds.) (1983) Concept Development and the Development of Word Meaning. Berlin/Heidelberg: Springer, 320-339.

Wannenmacher, W./Seiler, T. B. (1985a) Die Bedeutung verbaler Methoden für die Untersuchung von Wortbedeutungsentwicklung. In: Seiler, T. B. /Wannenmacher, W. (eds.) (1985) Begriffs- und Wortbedeutungsentwicklung: Theoretische, methodische und methodische Untersuchung. Berlin/Heidelberg: Springer, 193-210.

Warren, N. (ed.) (1978) Studies in Cross-cultural Psychology, Bd. I. London et al.: Academic Press.

Wegner, I. (1985) Frame-Theorie in der Lexikographie. Tübingen: Niemeyer.

Weinert, F. E./Waldmann, M. R. (1988) Wissensentwicklung und Wissenserwerb. In: Mandl, H./Spada, H. (eds.) (1988) Wissenspsychologie. München/Weinheim: Psychologie Verlag Union, 161-199.

Wettler, M. (1989) Wissensrepräsentation: Typen und Modelle. In: Bátori, I., S./Lenders, W./Putschke, W. (eds.) (1989) Computational Linguistics/Computer-Linguistik. An International Handbook on Computer Oriented Language Research and Application. Berlin/New York de: Gruyter, 317-336.

Wiegand, H. E. (1977) Nachdenken über Wörterbücher: Aktuelle Probleme. In: Drosdowski, G./Henne, H./Wiegand H. E. (eds.) (1977) Nachdenken über Wörterbücher. Mannheim: Bibliographisches Institut, 51-102.

Wiegand, H. E. (1985) Eine neue Auffassung der sog. lexikographischen Definition. In: Hyldgaard-Jensen, K./Zettersten, A. (eds.) (1985) Symponium on Lexicography: 2. Proceedings of the Second International Symposium on Lexicography May 26-17, 1984 at the University of Copenhagen. Tübingen: Niemeyer, 15-100.

Wiegand, H. E. (1988) Was ist eigentlich Fachlexikographie? Mit Hinweis zum Verhältnis von sprachlichem und enzyklopädischem Wissen. In: Munske, H. H./Polenz, P./Reichmann, O./Hildebrandt, R. (eds.) (1988) Deutscher Wortschatz. Berlin: de Gruyter, S. 729-790.

Wiegand, H. E./Wolski, W. (1980) Lexikalische Semantik, In: Lexikon der Germanistischen Linguistik. Tübingen: Niemeyer, 199-211.

Wierzbicka, A. (1985) Lexicography and Conceptual Analysis. Ann Arbor: Karoma Publishers.

Wierzbicka, A. (1996) Semantics: Primes and Universals. Oxford/New York: Oxford University Press.

Widdowson, H. G. (1978) Teaching Language as Communication. Oxford: Oxford University Press.

Wille, R./Zickwolff, M. (eds.) (1994) Begriffliche Wissensverarbeitung. Mannheim: B.I.-Wissenschaftsverlag.

Williams, M. E. (1987) Annual Review of Information Science and Technology, Bd. XXII. Amsterdam: Elsevier.

Winograd, T. (1975) Frame Representation and the Sclarative/Procedural Controversy. In: Bobrow, A. M./Collins, A. M. (eds.) (1975) Representation and Understanding. New York: Academic Press, 185-210.

Winston, M. E./Chaffin, R./Herrmann, D. J. (1987) A Taxonomy of Part-Whole Relations. In: Cognitive Science 11, 417-444.

Winston, P. H. (1975) The Psychology of Computer Vision. New York: McGraw-Hill.

Winter, E. O. (1974) Replacement as a Function of Repetition: a Study of Some of its Principal Features in the Clause Relation of Contemporary English. Dissertation, University of London.

Winter, E. O. (1976) Fundamentals of Information Structure: A Pilot Manual for further Development according to Student Need. Hartfield Poloytechnic.

Witten, I. H./Moffat, A./Bell, T. C. (1994) Managing Gigabytes: Compressing and Indexing Documents and Images. New York: Nostrand Rheinhold.

Wrobel, A. (2000) Textproduktion und Verfahren der Produktion schriftlicher Text. In: Brinker, K./Antos, G./Heinemann, W./Sager, S. F. (eds.) (2000) Text- und Gesprächslinguistik, Bd. I. Berlin/New York: de Gruyter, 459-472.

Wunderlich, D. (1976) Studien zur Sprechakttheorie. Frankfurt a. M.: Suhrkamp.

Yeh, C.-C. (2004) The Relationship of Cohesion and Coherence: A Constrative Study of English and Chienese. In: Journal of Laguage and Linguistics, 3, 2, 243-260.

Zampolli, A. (ed.) (1977) Linguistic Structures Processing. Amsterdam/New York: North-Holland.

Zimbardo, P. G. (1988) Psychologie. Berlin/Heidelberg: Springer.

Zimbardo, P. G./Gerring R. J. (1999) Psychologie. Berlin et al.: Springer.

Zimbardo, P. G./Hoppe-Graff, S./Keller, B./Engel, I. (1995) Psychologie. Berlin/Heidelberg: Springer.

Zimmermann, E. T. (1993) Zu Risken und Nebenwirkung von Bedeutungspostulaten. In: Linguistische Bericht 146, 263-282.

Zuell, C./Harkness, J./Hoffmeyer-Zlotnik, J. H. P. (eds.) (1996) ZUMA-Nachrichen Speziel: Textanalysis and Computers. Mannheim: ZUMA.