Computergestützte Textthema-Analyse

Magisterarbeit

Lennart Lopin

Institut für Germanistik
der
Universität Rostock

Rostock, den 02. Juni 2003

Betreuung: PD Dr. phil. habil. Bernd Skibitzki
und
Prof. Dr. rer. nat. habil. Andreas Heuer

Kurzfassung

Die Analyse eines Textes auf thematisch relevante Inhalte stellt in der Computerlinguistik eine große Herausforderung dar. “Intelligente” Programme, die ohne besondere Vorkenntnisse flexibel Texte zusammenfassen können, sind sowohl für kommerzielle Wissensverwaltung als auch für wissenschaftliche Recherchen von großer Bedeutung, insbesondere wenn es darum geht, in immer umfangreicheren Textkorpora gezielt Informationen aufzuspüren und kondensiert darzustellen. Ausgehend vom textlinguistischen Konzept der Isotopie untersucht die vorliegende Arbeit das sprachwissenschaftliche Modell der Topikketten (*lexical chains*) auf seine Anwendbarkeit bei der Textthemabestimmung. Den theoretischen Betrachtungen schließt sich eine Referenzimplementation an, die einen empirischen Nachweis für die Einsetzbarkeit von Topikketten bei der maschinellen Textzusammenfassung erbringen soll.

Abstract

Textual analysis of thematically relevant content remains a great challenge for computer linguistics. “Intelligent” programs, which can flexibly summarize texts without previous domain specific knowledge, are of great importance – both for commercial knowledge management systems as well as for scientific research – especially, because ever increasing electronic documents make systematic information assessment and condensation indispensable. Starting out from text-linguistic based concepts of “isotopy” the present work examines the linguistic model of lexical chains for its applicability on text extraction processes. These considerations are followed by a reference implementation which should produce an empiric proof of the feasibility of lexical chains for automatic summarization systems.

ACM CCS J.5 Linguistics,
H.3.1 Content Analysis and Indexing,
H.3.3 Information Search and Retrieval
General: automatic text summarization, discourse level summarization
Keywords: lexical chains, semantic database

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Teil I

Theoretische Überlegungen

1 Einleitung

Mit der stetig anwachsenden Menge von Informationen, die in kürzester Zeit gefiltert, analysiert und bewertet werden müssen, hat im Bereich der Computerlinguistik in den vergangenen Jahrzehnten das Forschungsgebiet der "Automatischen Textzusammenfassung" zunehmend an wissenschaftlicher und kommerzieller Bedeutung gewonnen.

Dazu beigetragen hat unter anderem die explosionsartige Vermehrung von Online-Dokumenten und elektronischen Publikationen sowie die Notwendigkeit von gezielten intelligenten Suchanfragen. Obwohl bereits in den fünfziger Jahren erste Versuche unternommen wurden, Texte auf ihren wesentlichen Kern zu reduzieren, erfuhr diese Forschungsrichtung – inzwischen als Teilgebiet des *information retrieval* anerkannt – seit den neunziger Jahren eine Renaissance. Begünstigt wurde diese Entwicklung neben neuen Forschungsergebnissen benachbarter Disziplinen (Kognitive Linguistik, KI-Forschung, Psychologie,...) auch dadurch, dass Rechen- und Speicherkapazitäten immer effizientere und anspruchsvollere Textanalyse-Algorithmen ermöglichten.

Neben der immer dringender empfundenen Notwendigkeit, in Wissenschaft und Wirtschaft spezielle Informationen aus unüberschaubaren Textmengen gezielt herauszudestillieren, spielt das korrekte Zusammenfassen eines Textes auch als Kriterium für den zukünftigen Einsatz künstlicher Intelligenz eine Rolle. Das dem Menschen eigene Unterscheidungsvermögen und seine Fähigkeit Wesentliches von Unwesentlichem zu trennen wird im Rahmen eines der Fernziele der Computerlinguistik – ein in natürlicher Sprache frei mit dem Menschen kommunizierendes System zu entwickeln – eine besondere Bedeutung zukommen.

Nichtsdestotrotz spielte dieser Teilbereich in der deutschen Forschung bislang eine eher untergeordnete Rolle. Auch das neuentfachte Interesse an der automatischen

Textzusammenfassung blieb bis auf wenige Ausnahmen auf den angelsächsischen Raum beschränkt, obgleich eigene vielversprechende theoretische Grundlagenforschungen älterer Natur zu diesem Themenkomplex längst existieren. Die vorliegende Arbeit versteht sich in diesem Sinne als Brückenschlag und beinhaltet den Versuch ein älteres, bisher unberücksichtigtes textlinguistisches Verfahren in Hinblick auf den Stand der internationalen Forschung aufzugreifen und auszuwerten. Innerhalb letzterer haben sich zudem seit Anfang der neunziger Jahre einige angelsächsische Wissenschaftler, auf der Suche nach einem verbesserten theoretischen Konzept, ebenfalls dem sprachwissenschaftlichen Ansatz der Topikketten (*lexical chains*) zugewandt, der das Fundament des hier zugrundeliegenden Ansatzes darstellen wird.

Nachdem Morris und Hirst[1] die Bedeutung der Topikketten für die Gewinnung des Textthemas heraustellten, forschen derzeit zwei Wissenschaftlerteams der Universitäten Delaware(USA)[2] und Ben Gurion(Israel)[3] in dieser, für die internationale Forschung relativ "jungen" Richtung.

Unabhängig vom späten Wiederentdecken des ursprünglich auf den französischen Sprachwissenschaftler Greimas zurückgehenden Begriffs der Isotopie hatte der deutsche Sprachwissenschaftler Erhard Agricola bereits in den frühen achtziger Jahren ein ansatzweise prozedurales System zur Textanalyse und Textthema-Ermittlung beschrieben. Unglücklicherweise hat sein strikt analytisches Modell bisher keinerlei technische Verifikation erfahren. Die gegenwärtige Arbeit soll diese Lücke schließen.

Der gewählte Ansatz konzentriert sich dabei auf die Suche nach dem Textthema und faßt diesen sprachwissenschaftlichen Begriff vom "Konzentrat eines Textes" im Sinne der Computerlinguistik als eine maximal kondensierte Textzusammenfassung auf. Darüber hinaus bietet das hier entwickelte Textanalysemodell zugleich Ausgangspunkte für eine Reihe weiterer Anwendungsgebiete innerhalb der Computerlinguistik (wie z.B. für die semantisch verfeinerte Suche innerhalb von Textkorpora). Dass das hier vorgestellte Modell auch für andere Anwendungsbereiche des *information retrieval* von Bedeutung sein könnte, sei an dieser Stelle lediglich angedeutet. Aus Platzgründen konnte im Verlauf der Arbeit auf diese interessanten "Querverbindungen" nicht weiter eingegangen werden.

[1] Morris, J.; Hirst, G. 1991. "Lexical cohesion computed by thesaural relations as an indicator of the structure of text". In: Computational Linguistics, Vol. 17, S. 21-42.

[2] H. Gregory Silber und Kathleen F. McCoy: An Efficient Text Summarizer Using Lexical Claims. In: Proceedings of the First International Conference on Natural Language Generation, INLG 2000, S. 268-271ff., Mitzpe Ramon, Israel, June, 2000.

[3] Regina Barzilay und Michael Eldahad: Using Lexical Chains for Text Summarization. In: Inderjeet Mani u. Mark T. Maybury (Hgg.): Advances in Automatic Text Summarization, Cambride 1999, S. 111ff.

2 Aufgabenstellung

Gegenstand dieser Magisterarbeit soll die theoretische Fundierung und Implementierung eines Algorithmus sein, der aus Beispieltexten deutscher Sprache das jeweilige Textthema ermittelt. Unter dem Textthema wird die semantisch maximal mögliche Komprimierungsrate (*compression rate*)[1] eines Textes verstanden. Ziel soll es sein, aus einem beliebigen[2] Text Kerninformationen zu extrahieren.

Im ersten Teil der Arbeit sollen die textlinguistischen Grundlagen zur Gewinnung eines solchen Textthemas untersucht werden. Dies wird im Wesentlichen in drei Schritten geschehen:

- Als Basis des anvisierten Algorithmus dient ein Textverständnis, dass auf dem Modell der Topikketten aufbaut. Die Grundlagen, Vorteile aber auch Schwierigkeiten dieses Ansatzes bei seiner Anwendung auf das automatische Textzusammenfassen sollen herausgestellt werden.
- Die Schwierigkeit der automatischen Erkennung und Aufstellung von Topikketten soll eingehend erläutert und eine Lösungsmöglichkeit vorgestellt werden.
- Bei der Analyse und Auswertung der Topikketten sollen insbesondere die Ergebnisse der Arbeiten von Erhard Agricola berücksichtigt und den notwendigen Modifikationen für eine anschließende Implementierung unterzogen werden.

Im zweiten Teil der Arbeit sollen die theoretischen Überlegungen durch die Implementierung eines eigenen Algorithmus verifiziert werden. Die zu entwickelnde Soft-

[1] Siehe (Man01, S. 3)

[2] bei entsprechender Datenbank auch aus einem nicht-deutschsprachigen Text

ware soll in der Programmiersprache C++ geschrieben werden. Als Hilfsmittel bei Entwicklung, kommt der Borland C++ Builder 5 zum Einsatz[3]. Die gewünschte Anwendung soll sich im Groben aus folgenden Software-Bestandteilen zusammensetzen:

- Ein geeigneter Part-of-speech-Tagger (PoS), der die notwendigen Lemmata aus dem Text zur weiteren Verarbeitung herausfiltern kann
- Eine semantische Datenbank als Grundlage für die Bildung von Topikketten
- Ein Analysemodul, dass die Lemmata zu Topikketten verknüpft
- Ein Gewichtungsmodul, dass nach einem bestimmten Schlüssel die wesentlichen Sätze aus dem Text auswählt.

Abschließend muß der implementierte Algorithmus einer ersten Evaluation unterzogen werden.

[3] Nachwievor zeichnen sich C und C++ durch hervorragende Compiler aus. Der Geschwindigkeitsvorteil, der durch die Verwendung dieser Programmiersprachen erwächst, ist in Hinsicht auf Analysen größerer Textkorpora nicht von der Hand zu weisen (selbst dann, wenn die Stringklassen in Java (java.lang.String) vielleicht bequemer Stringoperationen unterstützen). Der Borland Builder (http://www.borland.com/cbuilder/) wurde Visual C++ wegen seiner hochwertigen VCL-Klassenbibliothek vorgezogen.

3 Untersuchungsgegenstand

3.1 Das linguistische Grundverständis von Text

In der germanistischen Textlinguistik haben sich zwei Betrachtungsweisen bei der Klassifizierung des Phänomens *Text* etabliert: Auf der einen Seite versteht die frühe Textlinguistik, die sogenannte Textgrammatik, den Text als Bestandteil der *langue* im Sinne des Sprachsystems nach DE SAUSSURE. Dieser Auffassung zufolge, die selbst den Beginn der Erforschung textueller Erscheinungen markierte, ist die Konstitution von Texten gewissen Gesetzmäßigkeiten unterworfen. Zur Beschreibung dieser textuellen Regelmäßigkeiten bemühte die Textlinguistik Forschungsmodelle anderer Teildisziplinen und versuchte grammatische Strukturen im Text aufzuzeigen. Im Gegensatz zu dieser "Textgrammatik" untersucht die "textpragmatische Linguistik" seit den 70er Jahren verstärkt die kommunikative Struktur von Texten. Ihr geht es im Wesentlichen um die Betonung des situativen Charakters von Texten und um den Einfluß, den die Handlungsgebundenheit eines Textes auf seine Konstitution hat.

Aus diesen zwei unterschiedlichen wissenschaftlichen Positionen haben sich daher bei der Begriffsbestimmung des sprachlichen Phänomens "Text" zwei komplementäre Textbeschreibungen[1] herausgebildet:

Der Textgrammatik zufolge bildet der Satz die kleinste textuelle Einheit. Ein Text besteht dabei aus einer endlichen, wohlgeordneten Anzahl von Sätzen, die durch verschiedene syntaktische und semantische Beziehungen miteinander

[1] Vgl. hierzu z.B. (Bri92, S. 12-17), (MH02, S. 64-95). Die folgende Beschreibung ist stark verkürzt. Für einen hervorragenden Überblick über die unterschiedlichen Strömungen innerhalb der Textlinguistik siehe (WH91, S. 22ff.)

verknüpft sind. Das den Text zusammenhaltende Gewebe wird in einer Reihe von "Vertextungsmitteln", wie Konjunktionen, Pronomina, Proadverbien, Artikel, Deiktika u.a.m. verortet. Diese Beziehungen an der Textoberfläche wurden unter der Bezeichnung "Kohäsion" zusammengefasst. Kohärenz bezeichnet dahingegen die semantische Einheitlichkeit von Sätzen innerhalb eines Textes[2].

Die kommunikationsorientierte Textpragmatik versteht den Text vornehmlich als eine "Handlung". Die Bedeutung eines Textes, d.h. sowohl seine grammatische Verknüpfung als auch seine thematische Entfaltung, sind vom jeweiligen kommunikativen Kontext determiniert. Der pragmatisch kommunikative Zugriff weist dabei auf die unterschiedlichen Aussagen hin, die ein und derselbe Satz/Text in seinem jeweiligen Kontext annehmen kann. Für eine adäquate Textbestimmung führt die Textpragmatik deshalb weitere Textkriterien ein. Dazu zählen Intentionalität, Akzeptabilität, Informativität, Situationalität und Intertextualität[3].

Eine linguistische Definition von Text, die beide Erkenntnisse verbindet, lautet demnach wie folgt:

Satz 3.1 (Textdefinition)
"Unter Text versteht man eine wohlgeordnete, begrenzte Folge von sprachlichen Zeichen, die in sich kohäsiv und kohärent ist und die als Ganzes eine erkennbare kommunikative Funktion signalisiert"[4].

Für die vorliegende Untersuchung, d.h. bei dem Versuch einer analytischen *Inhalts*bestimmung, kommen allein textgrammatische Überlegungen in Betracht. Der Auffassung, dass ein adäquates textlinguistisches Gesamtmodell ohne eine Synthese beider Forschungsrichtungen unwahrscheinlich ist, soll damit nicht widersprochen werden[5]. Dennoch führt die Erweiterung des Textbeschreibungsmodells um die oben angeführten pragmatischen Kategorien zu einer Aufweichung des stringent strukturlinguistischen Modells und führt Kriterien ein, die sich einer objektiven Beschrei-

[2] Diese Unterscheidung wird nicht immer konsequent eingehalten. Oft werden unter dem Begriff der Kohärenz auch kohäsive Phänomene verstanden. Eindeutige Begriffsbestimmungen sind in der Textlinguistik selten. Daher sollte eine klare Abgrenzung, die sich zudem etabliert hat, genutzt werden. Die hier vorgenommene Verwendung des Begriffs schließt sich DE BEAUGRANDE und DRESSLER (RB81) sowie (Bri92, S. 21) an. Noch unschärfer ist die Begriffsabgrenzung dieser zwei Textphänomene allerdings in der Computerlinguistik, vgl. die Definitionen bei MANI (Man01, S. 92)

[3] In dieser Form neben anderen Modellen am umfassendsten, siehe (RB81)

[4] In Anlehnung an BRINKER, vgl. (Bri92, S. 17)

[5] (Bri92, S. 16)

bung entziehen. Eine automatische Bestimmung solcher kommunikativen Textfunktion stünde dabei vor nahezu unlösbaren Schwierigkeiten. Über Schlüsselwörter (sogenannte *cue phrases*) könnte ein Abschätzen z.B. der Intentionalität angestrebt werden. Danach müßte allerdings erst geklärt werden, wie die Ergebnisse einer solchen Bestandsaufnahme Auswirkung auf den Themengehalt eines Textes zeitigen. Ein im kommunikativen Zugriff pragmatisches Vorgehen ist aus der Sicht der Computerlinguistik zuerst einmal höchst "unpragmatisch".[6]

Im Mittelpunkt der vorliegenden Arbeit steht deshalb weniger die kommunikative Intention des Sprechers als die semantische Information des Textes. Das setzt voraus, dass dem Text ein semantischer Gehalt auch unabhängig von Produzent, Rezipient und Situation zugeschrieben werden kann. Obwohl diese Annahme höchst umstritten ist[7], können auf ihrer Basis durchaus hohe qualitative Ergebnisse erzielt werden[8].

Der maschinellen Erschließung des Textthemas muß also eine gewisse Vereinfachung und Abstraktion des Textes vorangehen. Im Rahmen des Kommunikationsmodells von Shannon und Weaver[9] soll der Text daher als Informationsträger (*code*) aufgefasst werden, der, mit einem Informationsgehalt (*message*) versehen, von einem Sender (*encoder*) an einen Empfänger (*decoder*) übermittelt wird. Im Weiteren soll davon ausgegangen werden, dass die sprachlichen Textmerkmale und ihre Relation untereinander (Kohäsion, als Bindeglied bei der sequentiellen Anordnung, und Kohärenz, als der semantischen Integrität aller Satzaussagen) ausreichende Rückschlüsse auf die wesentlichen, im Text eingebetteten Information erlauben.

Deshalb macht sich der hier gewählte Ansatz Eigenschaften der Textstruktur zunutze, um Informationen über das Textthema zu gewinnen; die Rolle der kommunikativ-pragmatischen Textfunktion wird vernachlässigt und die Textdefinition wie folgt vereinfacht:

Definition 3.2 (Eingeschränkte Textdefinition)
Unter "Text" versteht man eine wohlgeordnete, begrenzte Folge von sprachlichen Zeichen, die in sich kohäsiv und kohärent ist.

$$T = S_1, ..., S_n$$

[6] Inwieweit solche "weichen" Textbeschreibungskriterien beim Zusammenfassen des Textinhaltes wieder zur Geltung gelangen können und die Übersetzungsqualität verbessern helfen, siehe Karen Sparck Jones in: (IM99, S. 4)

[7] (Bri92, S. 55), sowie Kapitel 3.2

[8] Eine ausführliche Erörterung dieser Thematik findet sich in Kapitel 4.2, empirische Belege in Kapitel 7.1

[9] Claude E. Shannon und Warren Weaver: The Mathematical theory of Communication, University of Illinois Press, 1949 sowie (AL01, S. 174)

Gegen diese Auffassung vom Text als sequentielle Folge von Sätzen liegen mehrere Argumente vor. So wird zurecht von HEINEMANN festgestellt[10], dass es Fälle von "Texten" gibt, die aus nur einem Satz bestehen (Ezra Pounds Gedicht "Alba" z.B. oder japanische Haikus). Allerdings können diese Einzel-Satz-Texte ohne Schwierigkeiten als Spezialfälle von Texten aufgefasst werden.[11]

Als ein weiteres Argument gegen die Verwendung von Sätzen als Grundeinheiten von Texten könnten Ordnungsinstanzen transphrastischer Natur (wie z.B. Kapitel, Paragraph und Absatz) herangezogen werden. Dem praktischen Ansatz der vorliegenden Arbeit folgend, können diese satzübergreifenden Gliederungen als formale Strukturierung des Textes aufgefasst werden, die für den Inhalt desselben eine untergeordnete Rolle spielen[12]. Obwohl diese Anordnungen im Text im weitesten Sinne als kohäsives Mittel anerkannt werden könnten, da sie den semantischen Gehalt gewisser Textabschnitte hervorheben und formal verbinden, so läßt sich daraus kein Kriterium für den Textinhalt ableiten. Andererseits geht der Inhalt eines Textes generell nicht verloren, wenn die formale Gliederungsstruktur des Textes aufgehoben wird (wenn z.B. alle Absätze entfernt würden).

Ein weiterer Vorwurf gegen diese Art der Textdefinition lautet, dass sie den Text als fertige, in sich strukturierte Einheit kennzeichnet[13]. Dieser Vorwurf entpuppt sich, bei dem Versuch, die Grundlage für eine empirisch verifizierbare automatisierte Textanalyse zu entwickeln, als Vorteil. Ohne das Verständnis von einem fertigen, in sich sinnhaft gegliederten Untersuchungsgegenstand wäre ein Versuch, diesen über Algorithmen zu erschließen, ein unmögliches Unterfangen.

Somit muß die Loslösung von den am Kommunikationsprozess Beteiligten nicht als Nachteil, sondern als eine *Voraussetzung* der computergestützten Textanalyse verstanden werden. Diese Unabhängigkeit ist im Sinne einer möglichst weitreichenden Objektivität sogar wünschenswert. Da bei diesem Ansatz keine nicht-indizierten Phänomene berücksichtigt werden, können subjektive Auslegungen vermieden werden. Obwohl die hier gewählte prozedurale Textanalyse einen "generischen" (auf standardisierten Vorgaben beruhenden) Rezeptionsvorgang nachbildet[14], soll dieser jedoch eindeutig und umfassend beschrieben werden können.

---

[10] (MH02, S. 104)

[11] (Agr79, S. 32)

[12] Bei der Implementation sollen deshalb auch Überschriften als (Teil-)Sätze des Textes betrachtet werden.

[13] Dieser und folgende Einwände siehe (MH02, S. 68)

[14] Siehe insbesondere Kapitel 4.2

3.2 Die Theorie vom Informationskern

> "Man muß sich überhaupt darüber im klaren sein, dass es bei der textanalytischen Bestimmung des Themas (als Inhaltskern) keine "mechanische" Prozedur geben kann, die nach endlich vielen Schritten automatisch zur "richtigen" Themenformulierung führt."[15]

Neben den unterschiedlichen Textbestimmungsansätzen[16] divergieren in der Textlinguistik auch die Meinungen darüber, was als "Thema" eines Textes zu gelten hat. Während die Psycholinguistik z.B. das Thema als den "Fokus einer kommunikativen Interaktion"[17] versteht, bringen andere Linguisten[18] die Problemformulierung bei der Themabestimmung stärker in den Vordergrund. Größer noch werden die Widersprüche allerdings, wenn es darum geht, eine adäquate, alle textuellen Eigenschaft verbindende Thema-Analyse zu entwickeln.

Vor diesem Hintergrund scheint es kaum verwunderlich, dass Erhard Agricola in den frühen achtziger Jahren für seine Theorie von der Ermittlung des Informationskerns heftige Kritik geerntet hat[19]. In einem wesentlichen Punkt, nämlich dem strikt strukturalistisch-prozeduralen Vorgehen jedoch unterschied sich sein Modell von einigen anderen Beschreibungsansätzen[20]. Das Eingangszitat macht deutlich, wieso ausgerechnet Agricolas These vom Informationskern und seiner Entfaltung in weiten Teilen der Textlinguistik abgelehnt wurde. Zwar unterschied sie sich in der Grundannahme nicht wesentlich von van Dijks Vermutung, dass nämlich der thematische Kern einer Botschaft bei der Übermittlung von Emittent zu Rezipient in Gestalt eines Textes semantisch ausdifferenziert wird; aber selbst van Dijk beschrieb diese Thema-Entfaltung in relativ unscharfen "Makroregeln", die sich einer Automatisierung entzogen.[21]

Das Hauptargument gegen die Möglichkeit einer automatisierten, mithin algorithmischen Analyse der semantischen "Kerninformation" eines Textes lautet, dass die Bestimmung des Themas eine *Rezipiententätigkeit* voraussetzt[22]. Den Eindruck, den

[15] (Bri92, S. 55)
[16] Einen umfassenden Überblick bietet (MH02, S. 110)
[17] (Löt78, S. 18)
[18] (Löt78, S. 25)
[19] Diese kritischen Stimmen greift Lötscher in seinem Werk "Text und Thema" auf und gibt einen ausführlichen Überblick über Gegenstimmen zu Agricolas Position; siehe (Löt78, S. 35-46).
[20] So auch von van Dijks Makrostrukturenmodell, s.u.
[21] Ein Beispiel dafür liefern E. Gülich und W. Raible, zit. n. (Bri92, S. 56)
[22] "Der Leser [...] konstituiert im Lese-Vorgang den Textsinn [...] Dies vollzieht sich als ein spontaner, in unterschiedlichem Ausmaß unbewußter Prozeß", so z.B. Schutte, (Sch97, S. 158)

der Rezipient auf der Grundlage des vom Emittenten intendierten Bedeutungspotentials gewinnt, wird für ihn zum zentralen Thema eines Textes. Diese "Interpretation" des Stellenwertes einzelner im Text eingeschlossenen semantischer Werte wird also vor dem Hintergrund des Weltwissens und Erfahrungshorizontes eines jeden Rezipienten im Einzelnen subjektiv anders aussehen. Dieses Phänomen hat die Hermeneutik mit dem Begriff der *hermeneutischen Differenz* umfassend beschrieben (siehe Abbildung 3.1)[23].

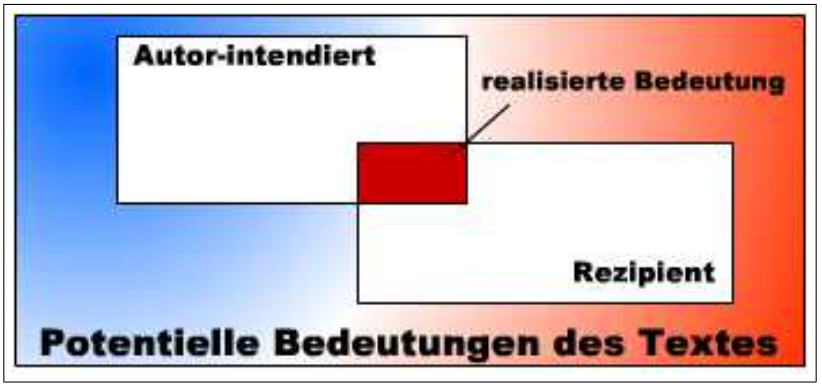


Abbildung 3.1: Hermeneutische Differenz

Ein Voraussetzen von Weltwissen und individueller Lebenserfahrung als wesentliche Komponenten bei der Bedeutungskonstitution eines Textes führt dazu, dass ein objektiver Zugriff auf Inhalt und Thema von vornherein abgelehnt wird. Und doch bildet die Abgeschlossenheit eines Textes und somit sein begrenztes, in der Struktur des Textes verankertes Bedeutungsspektrum die Grundlage für eine subjektive Rezipiententätigkeit.

Der Text ist daher keine chaotische Informationsmenge, in die erst durch den Rezipienten vollständige Ordnung einkehrt. Wie aus der Textdefinition 3.1 ersichtlich wurde, herrscht im Text eine gewisse Ordnung vor. Dies entspricht eindeutig seiner Rolle als Träger der Information im Sinne des Kommunikationsmodels. Weil die Satzbedeutungen eben in Hinblick auf eine *bestimmte* Textaussage vom Emittenten evoziert worden sind, muß sich die intendierte Textbedeutung an der Textoberfläche in Kohäsion und Kohärenz bemerkbar machen.

Unsere Vermutung lautet daher, dass unabhängig vom Emittenten und Rezipienten die Ordnung der im Text enthaltenen Einzelbedeutungen etwas über die tatsächliche bzw. am wahrscheinlichsten zu aktualisierende Textthematik aussagen kann.

So wie also die Satzbedeutung dadurch entsteht, dass, dem Kontext entsprechend, Sememe der Einzellexeme ausgewählt und zu einer Proposition integriert werden,

[23] Eine weiterführende Darstellung liefert (Sch97, S. 23)

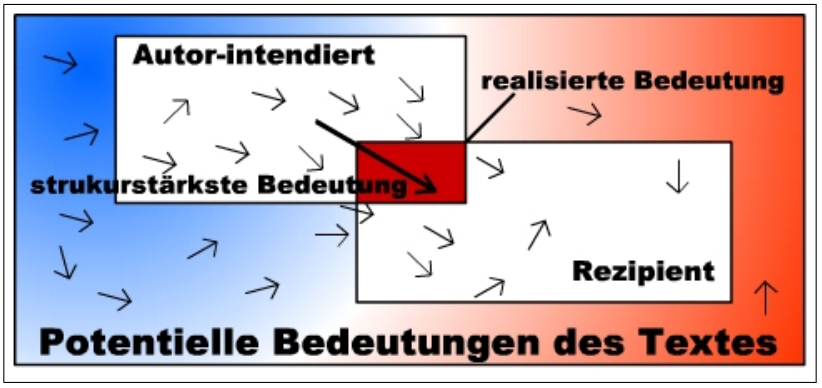


Abbildung 3.2: Rezipientenunabhängige Bedeutungsstruktur

so kann auch das Textthema als eine Integration der Satzpropositionen aufgefaßt werden.

Dieser Auswahl- und Integrationsprozeß ist eben jene schwer fassbare Arbeit, die der Rezipient beim Aufnehmen eines Textes nahezu unbewusst verfolgt. Auswahlkriterien für die Aktualisierung und das Verwerfen bestimmter Semembedeutungen mögen zu einem gewissen Grade seinem bisherigen Erfahrungsschatz (im Umgang mit Texten und Sachverhalten der wahrnehmbaren Welt) geschuldet sein. Doch auch er kann nicht beliebige Bedeutungen auswählen. Dies verhindert die Konventionalisierung der sprachlichen Zeichen; der enge Rahmen aller möglichen Sememe pro Lexem bildet dabei die untere Grenze der Bedeutungsvariabilität.

Unser Versuch muß nun darin bestehen, diesen Prozeß so weit wie möglich zu objektivieren. Sollte dies auf der Grundlage der Oberflächenstruktur des Textes gelingen, könnte auch bei Ausschaltung einer subjektiven Bewertungskomponente ein "generischer" automatisierter Auswahlprozeß die semantisch-thematischen Komponenten des Textes erfassen.

Wie aber kann die Oberflächenstruktur eines Textes auf den Inhalt hinweisen? Eine Antwort liegt im hypothetisch formulierten Informationskern begründet, der jeder Textproduktion zugrundeliegen soll: Agricola zufolge findet der gesamte Kommunikationsprozeß im Schwerfeld dieses Textkerns, des Themas, statt. Dabei wird ein Leit- oder Grundgedanke in "linear-sequentieller" Abwicklung in Form des Textinhaltes ausgebreitet (die sich in der Oberflächenstruktur widerspiegelt). Die abstraktsemantische Konzentration dieses Leitgedankens, das Thema, enthält dabei als Abbild eines komplexen Sachverhaltes bereits alle Beziehungen und Prädikate, die dann in Folge daraus abgeleitet und im Text versprachlicht werden.

Damit sind die eingangs festgehaltenen Oberflächenstrukturen Kohäsion und Kohärenz Ableitungen aus der Progression eines thematischen Bezuges, die sich, wie

es AGRICOLA ausdrückt, "mittelbar" in kohäsiven und "eigentlich"[24] in kohärenten Beziehungen als komplexe Verflechtung der Einzeltexteme manifestiert.[25]

Die Kritik an diesem Ansatz richtet sich im wesentlichen stets gegen die angestrebte "Objektivität"[26]. Unter den Prämissen der hermeneutischen Differenz, muß eine Vorgehensweise zur Bestimmung des am "wahrscheinlichsten" geltenden Themas ermittelt werden, wobei sich dieses auf der Grundlage einer am Text aufzeigbaren semantischen Indifferenz darstellt und in einer strukturstärksten Bedeutung manifestiert.

Die automatische Textthema-Analyse muß dabei aus der Textstruktur eine Themenstruktur aufbauen und daraus wiederum das Thema ableiten, das im Mittelpunkt der geordneten Textstruktur steht. Dies wird mit Sicherheit die weitreichende Forderung nach "Objektivität", d.h. nach einer Überprüfbarkeit[27] der Textanalyse, so weit erfüllen, wie es bei einem derart polysemen Untersuchungsgegenstand wie dem Text überhaupt möglich ist. Dabei muß es sich nicht um das "gewünschte" Thema des Textes, im Sinne der Absicht des Emittenten bzw. der Erwartung der Rezipienten handeln, sondern um die aus der Anordnung und Verflechtung der sprachlichen Zeichen ableitbare Thematik. Ob ein Autor diese strukturstärkste Bedeutung bei der Textproduktion angestrebt hat, oder die Rezeptionskompetenz eines Leser in der Lage ist, eine solche aufzuspüren, liegt außerhalb der hier zu behandelnden struktursemantisch orientierten Textanalyse. In jedem Fall führt sie zu einem Extraktionsergebnis, dass menschlichen Textzusammenfassungen sehr nahe kommt und zu erstaunlichen Ergebnissen führen kann[28].

Das sprachliche Phänomen Text ist somit, unter dem Vorzeichen einer automatisierten Analyse, klar definiert worden. Kohäsion und Kohärenz haben sich als zentrale Merkmale zur Textbeschreibung herausgestellt. Der Anspruch der Informationskernhypothese das Textthema eindeutig zu beschreiben fand, unter Berücksichtigung der hermeneutischen Differenz, durch das Ableitbarkeitsprinzip[29] eine Bestätigung.

---

[24] (Agr79, S. 33)

[25] Zum Ableitbarkeitsprinzip siehe (Bri92, S. 56)

[26] vgl. (Bri92, S. 56) u. (Löt78, S. 41f.)

[27] und beliebigen Wiederholbarkeit auf der Grundlage standardisierter Kriterien. Kapitel 4.2, 4.4 und 6.2 werden näher darauf eingehen, auf welchen Kriterien eine Objektivität bei der computergestützten Thema-Analyse erreicht werden kann.

[28] Siehe Kapitel 7.1

[29] Ableitung des Textes aus einem thematischen Kern, siehe auch (Bri92, S. 22, 56, 59)

4 Vom Text zum Thema

4.1 Das Topikkettenmodell

Zwischen der Oberflächenstruktur eines Textes mit ihren syntaktischen Verflechtungen und den inhaltlichen, thematischen Aspekten besteht eine enge Verbindung, die sich im Merkmal der Kohärenz offenbart.

Um von der Oberfläche zum thematischen Gerüst eines Textes vorzudringen und von einer lexikalischen oder syntaktischen zu einer semantischen Makrostruktur aufzusteigen, liegt es nahe, sich die Ergebnisse des Isotopiemodells[1] nach Greimas zunutze zu machen[2]. Das Verständnis von der Isotopie eines Textes legte den Grundstein für die Überwindung einer ausschließlich syntaktischen Betrachtungsweise und führte zu einem Miteinbeziehen semasiologischer Faktoren bei der Textanalyse.

Worauf kann ein semantischer, d.h. inhaltlicher Zusammenhang von Texten gestützt sein? Welches System steckt hinter der Verteilung kohärenter Erscheinungen im Satz? Für die Beantwortung dieser Fragen formulierte der französische Sprachwissenschaftler Algirdas Julien Greimas den Textbegriff neu. In seinen Augen ist ein Text vornehmlich ein semantisches Gebilde und erst in zweiter Linie ein syntaktisches Konstrukt. Aus diesem Grunde ist es besser, dass man bei der Textdefinition die Bausteine des Textes nicht direkt mit dessen Einzelsätzen gleichsetzt. Deshalb war in Definition 3.2 generell von sprachlichen Zeichen die Rede, die den Text auf-

[1] Die Isotopie zählt neben der thematischen Progression zu den fundiertesten Vertretern semantischer Textbeschreibungsmodelle, siehe (MH02, S. 72)

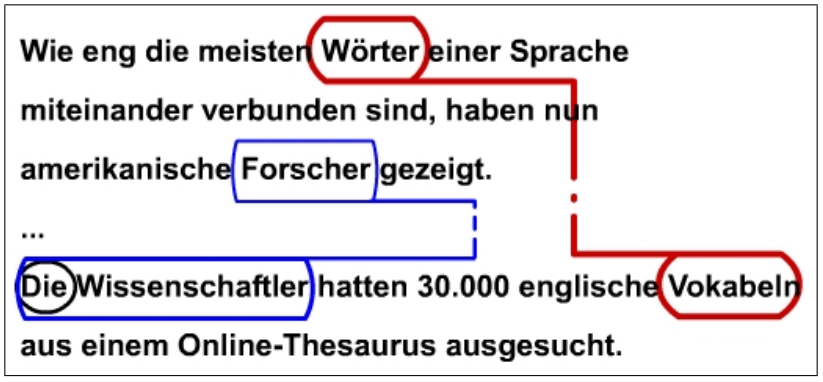
[2] Grundlagenwerke dieser semantischen Sprachtheorie bilden Bernhard Pottiers "Linguistique général: théorie et description. Paris, 1974 und Algirdas Julien Greimas' "Sémantique structurale. Paris, 1966

bauen. Greimas versteht darunter hauptsächlich Lexeme, für die der Satz einen noch näher zu erläuternden "Verknüpfungshintergrund" darstellt.

Isotopie ist demnach die Verkettung von Lexemen auf semantischer Ebene. Grundlage für diese semantische Äquivalenz zwischen zwei oder mehreren Segmenten eines Textes liefern Lexeme, die über Semrekurrenz und Referenzidentität in Relation stehen und eine sogenannte Isotopieebene bzw. Topikkette aufmachen.[3]

Der Aufbau einer Topikkette erfolgt gemäß dem Grad der semantischen Nähe zweier Lexeme. Dabei spielen die Seme eines Lexems eine entscheidende Rolle. Sie entscheiden darüber, ob ein bestimmtes Lexem an einer Topikkette überhaupt teilnehmen kann, und innerhalb einer Isotopiekette sorgen sie dafür, dass ein ganz bestimmtes Sememen des Lexems in den Vordergrund rückt. Ein Beispiel soll diesen Sachverhalt[4] verdeutlichen:[5]

Abbildung 4.1: Topikketten



Wie aus einem etwas ausführlicheren Beispiel[6] ersichtlich wird, entspricht eine solche Topikkette einer thematischen Achse des Textes. Diese Trägerfunktion erfüllt sie durch die Einengung der Bedeutungsmenge aller am Text beteiligten Einzellexeme. Der Satz bietet dabei den Rahmen für die Querverbindungen zwischen den einzelnen Topikketten.

[3] In Anlehnung an (Hei00, S. 54-59)
[4] Wobei $Tk_1(Wort \simeq Vokabel)$ und $Tk_2(Forscher \simeq Wissenschaftler)$
[5] Zur Semrekurrenz siehe (Hei00, S. 54-59) und (MH02, S. 72)
[6] Siehe Anhang Textbeispiele, Tabelle 17.1, S. 81

Dabei lassen sich in erster Instanz drei Spielarten von Semrekurrenzen unterscheiden[7]:

- Die direkte Lexemrepitition, oder auch Reiteration, bei der ein und dasselbe Lexem im Text wiederholt wird.
- Semrekurrenz durch variierende Wiederholung auf der Grundlage von lexikalischen Relationen wie Hyperonymie, Hyponymie, Synonymie, Antonymie, etc.
- Die grammatikalische Substitution, bei der z.B. ein Substantiv durch Artikel oder Pronomen ersetzt wird[8].

Auf dem Fundament des Isotopiemodells läßt sich somit ein ganze Anzahl von textuellen Phänomenen erklären. Kohärenz ist in diesem Sinne ein Ergebnis der satzübergreifenden Topikketten, die den Text in unterschiedlicher Länge durchziehen. Diese Kohärenz drückt sich wiederum in den einzelnen kohäsiven Textoberflächenmerkmalen aus.

Die "Zuarbeit", die die Kohäsion leistet, zeigt aber auch, dass sie in Bezug auf die Textthematik eine der Kohärenz untergeordnete Rolle spielt. In der Tat war die simple Aneinanderreihung von koreferierenden Substantiven (Reiteration) und ihre quantitative Analyse eine der ersten Methoden zur automatischen Textzusammenfassung.[9]

Außerdem weist das Isotopiemodell den Lexemen als Bausteinen des Textes eine zentrale Rolle zu. Als "Textatome" binden deren Seme die Topikketten zu thematischen Einheiten und öffnen verschiedene semantische Ebenen, die von Satz zu Satz einander gegenübergestellt und durch ihre verschiedenen Relationen zueinander die Thema-Entfaltung widerspiegeln. Vor diesem Hintergrund fungiert ein Satz als Topikkettenumgebung. Seine Funktion besteht in der Zuordnung zweier oder mehrerer Topikketten zueinander.

Wenn ein Satz als relationaler Hintergrund für verschiedene ihn durchkreuzende Topikketten aufgefaßt wird, so kommen für die Lexeme, die an den Topikketten

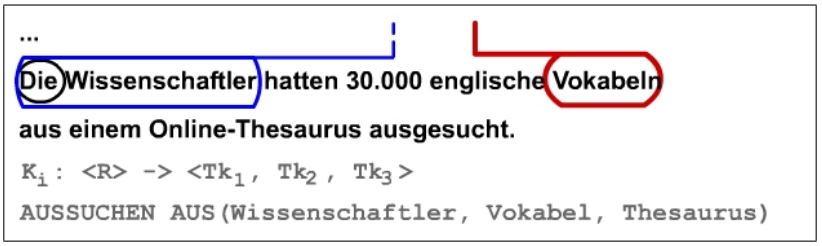
[7] so (MH02, S. 72); von den drei angeführten Topikketten-Varianten werden im Verlauf der Implementation nur die ersten beiden zur Bildung von Topikketten herangezogen, da insbesondere die "Anaphernresolution" einen aufwendigen Analyseapparat nachsichziehen würde, der in keinem Verhältnis zum zusätzlich erlangten Gewinn stünde.

[8] Vgl. dazu Abbildung 4.2, wo der schwarzumrandete Artikel "die" ein innertextuelles Objekt referenziert und damit eine Zuordnung zur Topikkette erleichtert

[9] Edmundsons *key word*-Extraktion, vgl. (Man01, S. 48)

beteiligt sind im Wesentlichen Autosemantika in Frage. Die (syntaktische) Subjekt–Objekt–Prädikat Struktur eines Satzes wiederum bietet dabei die Grundlage für eine Manifestation der einzelnen (lexikalischen) Kettenglieder auf der Satzebene. Im syntaktischen Verhältnis zwischen Subjekt und Objekt sowie der entsprechenden Prädikation findet die Zuordnung der Topikketten untereinander statt.

Abbildung 4.2: Satzrelation



Das Isotopiemodell liefert also eine feste Grundlage, um eine Fülle von Textmerkmalen auf ihre Funktion für die Textbedeutung hin verstehen und einordnen zu können. Außerdem ist sein Charakter äußerst praktikabler Art, ist doch der Ausgangsbaustein, das Lexem, mittels Wortform-Extraktion aus Texten einfacher zu handhaben als beispielsweise Satzganzheiten. Bevor das genauere Vorgehen für die Gewinnung eines Themas auf der Grundlage der Textisotopie besprochen werden kann, müssen etwaige Beschränkungen dieses Modells untersucht werden.

So hat sich im Laufe der Zeit, seit der Einführung des Isotopiemodells in der Linguistik, immer deutlicher gezeigt, dass man bei der Bildung von Topikketten auf ein Referenzproblem stößt. Topikketten können z.B. an Sachverhalte anknüpfen, die nicht explizit im Text ausgedrückt werden, oder Komposita setzen sich aus Lexemen zusammen, die zwei unterschiedlichen Topikketten gleichzeitig zugeordnet werden könnten. In der Denotatsreferenz und Referenzambiguität liegt ein großes Problem bei der Verwendung von Topikketten in der Textanalyse. Die Überschneidungen von Topikketten können zumeist erst vor dem Hintergrund der gesamten Topikkettenstruktur des Textes aufgelöst werden.

Manchmal verweisen Topikketten auf Zusammenhänge in der “realen” Welt. Die Semrekurrenz wird dann nicht durch einfache Synonymie oder hyper- bzw. hyponymische Verhältnisse bestimmt, sondern von außersprachlichen, vielleicht sogar okkasionellen Gegebenheiten. Die Pragmalinguistik findet hier erneut ihre Bestätigung, indem der Kontext der Kommunikationshandlung konstituierend auf den Text einwirkt.[10]

[10] (Hei00, S. 58)

Auch AGRICOLA geht auf dieses "nicht zu unterschätzende" Problem ein und versteht unter der sogenannten "Koreferenz":

> "den Bezug auf ein Wirklichkeitsmodell (und damit indirekt auf die objektive Realität) im Kommunikationsakt, dessen Herstellung Sprecher und Hörer vollziehen."[11]

Bei der Festlegung, ob zwei, isoliert betrachtet, äquivalente Sememe zu ein und derselben (Indikatoren) Isotopiekette gehören oder zu zwei unterschiedlichen (Alteratoren), spielen nach Agricola vor allem die Pronomina, Artikel und Zahlwörter eine Rolle.

Dieses Koreferenzproblem führt in der Tat zu technischen Schwierigkeiten. BARZILAY und ELHADAD schlagen als eine mögliche Antwort zu diesem Problem eine Gesamtanalyse aller möglichen Isotopieketten vor. Bei ihrem Versuch, einen geeigneten Algorithmus zu definieren, der die Aufnahme eines Lexems in eine bestimmte Topikkette nicht dem puren Zufall überläßt, berechnen sie nebeneinander alle durch die Semrekurrenz entsprechender Lexeme aufzustellenden möglichen Topikketten und gewichten diese jeweils nach dem Auftreten von Reiterationen, Synonymen, Antonymen und Hyperonymen in absteigender Folge. Ein Schwellenwert sorgt dafür, dass ab einem beliebig festzulegenden Wert nur die stärkste der gerade nebeneinander entwickelten Topikketten bestehen bleibt.

> "Under the assumption that the text is *cohesive* we define the best interpretation as the one with the most connections [...] We define the score of an interpretation as the sum of its chain scores. A chain score is determined by the number and weight of the relations between chain members."[12]

Falls man die Relationen zwischen den Lexemen erfassen kann, ist somit eine Disambiguierung der Topikketten auch auf diesem (innertextuellen) Wege möglich. dass sich der "Greimas'sche Begriff der Isotopie nur schwer zu fassen und zu operationalisieren"[13] läßt, kann somit, wie BARZILAY und ELHADADs Implementierungserfolge beweisen, nicht bestätigt werden.

Ein weiteres Problem des Isotopiekonzeptes hat zuerst WOLFGANG HEINEMANN zur Sprache gebracht. Er weist darauf hin, dass die Isotopie und damit die lexikalische

[11] (Agr79, S. 47)
[12] (IM99, S. 114)
[13] zit. n. (Hei00, S. 58)

Semantik keine hinreichende Bedingung für das Textverstehen liefern. HEINEMANN führt zum einen den Sonderfall eines Ein-Satz-Textes[14] an und meint, dass sich in diesem – trotz eines gewissen Zusammenhangs – keine Topikketten ausmachen lassen:

> Silberne Wasser brausten, süße Waldvögel zwitscherten, die Herdenglöckchen läuteten, die mannigfaltigen grünen Bäume wurden von der Sonne goldig angestrahlt. (H. Heine, Die Harzreise)

Bei der detaillierteren Beschreibung der Topikkettenaufbaus wird sich herausstellen, dass die einzelnen Seme der Lexeme dieses Beispielsatzes sich über die Kategorie "Natur" schneiden. Das "Superthema", dass HEINEMANN in diesem Satz entdeckt, leitet sich im Grunde vom Anklingen der sich überschneidenden Seme her. Folgen weitere Sätze, so wird entweder diese einem "Superthema" nahe kommende Topikkette weiterentwickelt, oder einzelne Lexeme dieses Satzes bilden mit Folgelexemen eigenständige, weil stärker korrelierende Ketten, die vor dem Satzhintergrund mit den übrigen Topikketten interagieren. Auch in diesem Beispiel können isotope Erscheinungen ausgemacht werden[15].

Der zweite Gegenbeweis den HEINEMANN heranzieht, um die Grenzen der Topikketten bei der Textanalyse aufzuzeigen, wendet sich gegen die Argumentation, dass sich Texte allein aus dem Merkmal der Semrekurrenz ableiten. Für diesen Zweck führt er das folgende, bereits klassisch gewordene Beispiel an:[16]

> "Es gibt niemanden, den ihr Gesang nicht fortreißt. Unsere Sängerin heißt Josephine. Gesang ist ein Wort mit fünf Buchstaben. Sängerinnen machen viele Worte."[17]

Obwohl die Topikkette "Gesang" eindeutig auszumachen ist, stehen die Sätze in keinem (bekannten) Sinnzusammenhang. Hier setzt HEINEMANNs Kritik an. Allerdings kommt man nicht umhin festzustellen, dass eine solche Topikkette wie in diesem Beispiel für einen gewissen Zusammenhang sorgt. Damit entsteht der Eindruck einer beabsichtigten Zusammenstellung der Sätze. Die Rätselhaftigkeit, die eine solche

[14] (WH91, S. 39) und (MH02, S. 74)

[15] Der Isotopiebegriff wird hier und in Folge erweitert aufgefasst, indem die Semrekurrenz als das entscheidende Merkmal für die Topikkettenbildung ausgenutzt wird. Daher schließen Topikketten nach dem Modell AGRICOLAS auch "Kontiguitätsketten" (zu diesem Begriff siehe BRINKER, (Bri92, S. 36)) mit ein.

[16] Siehe auch http://computerphilologie.uni-muenchen.de/jg01/tiutenko.html, (25.05.2003)

[17] zit. n. (WH91, S. 40)

Topikkette mit ihren Teilgliedern vor dem Hintergrund des Weltwissens eines Rezipienten hervorrufen kann, läßt sie nicht zwangsläufig "unsinnig" erscheinen. Einen anderen Anspruch als den, Sinnstrukturen im Text aufzudecken, würde dem Isotopiemodell nicht gerecht werden und würde bedeuten den Rezipienten mit seinem gesamten Erfahrungshorizont in die Analyse einzubeziehen. Dagegen spricht der eingangs erläuterte[18] und in dieser Untersuchung angestrebte Objektivierbarkeitsgrad, der in der Textanalyse erreicht werden soll.

Auf die Widersprüchlichkeit dieser letzten beiden Einschränkungsversuche muß deshalb hingewiesen werden, weil ein näherer Blick zeigt, dass die Semrekurrenz nur die Grundlage für einen vom *Rezipienten* vorgenommenen mentalen Verknüpfungsvorgang[19] liefert. Selbst in solchen extremen Sonderfällen, die sozusagen die obere und untere Schranke der aus Semrekurrenzen überhaupt möglichen Topikketten (und damit sinnhaltigen Texten) darstellen, macht sich der Emittent diese *Dekodierungsversuche* des Rezipienten zunutze, um einen Sinngehalt zu übermitteln. Indirekt hat Heinemann damit eine weitere Bestätigung für die Mächtigkeit des Topikkettenmodells geliefert.

4.2 Textaktanten und Informationsknoten

Die Umformungsoperationen, die bei der Dekodierung des ausgebreiteten Textes zur Gewinnung eines Informationskernes führen, sind von zentraler Bedeutung für die prozedurale Beschreibung der Textthema-Analyse. Es muß ein Weg gefunden werden, um von diesem eigentlich mentalen Vorgang zu abstrahieren. Mithilfe der Isotopie und unter der Voraussetzung von Kohäsion und Kohärenz eines Textes stellt Agricola einen möglichen Analyseansatz exemplarisch vor[20].

Dabei postuliert er, dass es einige wenige Haupt-Isotopieketten geben muß, die in der Länge und Anzahl von Lexemen mit Abstand vor allen anderen Topikketten rangieren. Er verweist in diesem Zusammenhang auf die elementare kognitive Funktion des Präzidierens, bei der Sachverhalte aus der objektiven Welt im Bewußtsein durch Objekte und ihre entsprechenden Relationen (Verben) nachempfunden werden. Obzwar in einem Text viele unterschiedliche Sachverhaltsrelationen angesprochen werden können, muß es eine ausgewählte Reihe von Objekten geben, um derenwillen das Textaufkommen produziert wurde und die im Rahmen des Kommunikationsobjektes übermittelt werden sollen.

[18] Vgl. die Überlegungen zur Hermeneutischen Differenz S. 20

[19] Siehe Kapitel 4.4

[20] Und zwar an zwei Texten: Einmal an einer Zeitungsmeldung (Agr79, S. 43-71), danach an einem literarischen Text (Agr79, S. 72-94)

Wie aber können diese für das Thema des Textes wesentlichen Handlungsträger aus der Menge der behandelten Objekte herausgefiltert werden? Agricola geht davon aus, dass sie sich in den Lexemen (Sememen) der Hauptisotopieketten wiederspiegeln, solange der Text nicht selbst bereits ein Kondensat darstellt und die allgemeine Kürze der Topikketten eine Identifikation der gewichtigeren erschwert. Agricola stellt unter diesen Voraussetzungen folgende Behauptung auf:

> "Diejenigen Textausschnitte, in denen unterschiedliche Textaktanten als Elemente unterschiedlicher Haupt-Isotopieketten miteinander in Aktion treten, d.h. die Teilstrukturen, wo sie untereinander im Sinne der Syntax semantischer Elementarstrukturen oder logischer Propositionen über ein gemeinsames Prädikat verknüpft sind, bilden die markanten und unverzichtbaren Segmente, auf denen die zusammenhängende und sinnvolle inhaltliche Textprogression sich gewissermaßen wie auf Brückenpfeiler stützt und auf denen die *Durchgängigkeit* des Textes beruht."[21]

Textaktanten sind Abstraktionen der in den Hauptketten in einer bestimmten Bedeutung monosemierten Lexeme. Diese Aktanten sind damit von "synonymisch-paraphrasischen" Einheiten abstrahierte, verallgemeinerte Bedeutungen der quantitativ längsten Haupttopikketten, die für den referenzierten Sachverhalt stehen, um den das Thema eines Textes kreist.

Sobald also Textaktanten zweier oder mehrerer Haupttopikketten in einem Textem miteinander in Verbindung treten (vermittels einer Prädikation in eine Relation gebracht werden), muß es sich bei diesem Textausschnitt (Segment) um einen wesentlichen Bestandteil des Textes handeln.

Diese schlichte Tatsache legt den ersten Grundstein für eine erfolgreiche – weil objektivierbare – Textthema-Analyse und bildet den wichtigsten Ansatzpunkt für eine prozedurale Textthema-Ermittlung auf der Basis von Topikketten.

Die ausgezeichneten Textsegmente, sogenannte "Interaktionsknoten"[22], die sich aus zwei oder mehreren Textaktanten und einem Prädikat (dem "Relator") zusammensetzen, bilden damit die Elementareinheiten für eine Ermittlung der relevanten Themenstruktur. Ein Sonderfall bilden Sätze, in denen ein Textaktant allein über ein Verb spezifiziert wird. Es muß eine Möglichkeit gefunden werden auch solche Sätze entsprechend ihrer Wertigkeit für das Thema zu berücksichtigen.

[21] (Agr79, S. 42)
[22] (Agr79, S. 42)

Die Hypothese von einigen wenigen Haupttopikketten, die in ihrer Lexemanzahl signifikant vor den restlichen Ketten eines Textes rangieren, läßt vermuten, dass ein Text, der nur noch aus gleichgewichtigen, weil gleich langen Topikketten besteht, bereits einer Kondensation unterzogen wurde.

Ein Textthema darf somit im Idealfall nur gleich lange (möglichst kurze, im Idealfall "einstellige") Topikketten aufweisen. Lassen sich quantitative Unterschiede feststellen, so ist der Text noch weiter reduzierbar.

Diese Ergebnisse sollen in folgendem Satz noch einmal verdeutlicht werden:

Satz 4.1 (Satz nach Agricola)
Die über die Semrekurrenz der einzelnen Textem-Lexeme aufzustellenden Topikketten zerfallen in Haupt- und Nebentopikketten, die sich in der Anzahl ihrer Lexemglieder quantitativ signifikant unterscheiden. Textsegmente, in denen zwei oder mehrere Textaktanten über eine Prädikation in Relation gebracht werden, heißen Interaktionsknoten. Aus ihnen setzt sich das Thema eines Textes zusammen.

Eine direkte Anwendung dieses Satzes auf den Text führt zu einer ersten Annäherung an das Thema, indem, nach Aufstellung der Haupttopikketten, die längste aller Haupttopikketten über die Schnittmenge der zahlenmäßig nachfolgenden Haupttopikketten eine Anzahl von Textemen herauskristallisiert, die dem obigen Satz entsprechend, die Bedingung für Interaktionsknoten sowie eine thematragende Funktion erfüllen. Für den Beispieltext[23] wären dies:

$HT_{k3} \cap HT_{k6}$	$\cap HT_{k1}$
T_2	–
T_4	√
T_9	–
T_{10}	–
T_{12}	√

Tabelle 4.1: Texteme, geordnet nach der Schnittmenge der 3 Haupttopikketten T_{k3}, T_{k6} und T_{k1}

Wie aus der Tabelle zu ersehen, führt das Abgleichen der drei Haupttopikketten mit den Sätzen des Beispieltextes zu einer Auszeichnung der Texteme $T_4 und T_1 2$. Durch diese beiden Sätze laufen damit die den Text maßgeblich bestimmenden Topikketten – ein Zeichen für den Anteil dieser beiden Sätze am Gesamtthema des Textes.

Unter den Prämissen des Satzes nach Agricola lassen sich auch einige Berechnungen zur Gewichtung von Textemen begründen, die von Buchbinder und Rozanov

[23] Siehe Anhang A, S. 79 sowie S. 82

bereits Mitte der siebziger Jahre vorgestellt wurden[24]. Mit ihrer Hilfe kann die Suche nach den thematisch relevanten Textsegmenten verfeinert werden. Diese Koeffizienten, die neben der Schnittmengenbildung zur Textthemabestimmung herangezogen werden könnten[25], sollen an dieser Stelle eingehender behandelt werden, vor allem deshalb, weil sie scheinbar auch der angelsächsischen Fachliteratur unbekannt sind.

Vermittels des Koeffizienten der **semantischen Belastung** K_s kann die Wertigkeit eines beliebigen Textems im Verhältnis zum Isotopiegefüge des Textes näher bestimmt werden:

$$K_s = \frac{L}{N}$$

Dabei steht L für die Anzahl der Isotopieverbindungen der "Stützwörter" des betreffenden Textems und N für die Anzahl der Lexemglieder der jeweiligen Topikketten, einschließlich des Stützwortes des gegebenen Textems.[26]

Die unterschiedlichen thematischen Verdichtungszentren lassen sich über den **lokalen Koeffizienten** K_l voneinander abgrenzen. Damit werden inhaltliche Abgrenzungen und Segmentierungen im Text deutlich. Diese thematischen Schwerpunktbildung liefert in den meisten Fällen ein genaueres Bild von der thematischen Struktur eines Textes, als K_s[27]. Der lokale Koeffizient zwischen zwei beliebigen Textemen wird wie folgt ermittelt:

$$K_l = \frac{R}{N}$$

Hier steht R für die Anzahl aller Isotopierelationen zwischen den Lexemen zweier beliebiger Texteme eines Textes und N für die Menge der Stützwörter (Autosemantika) in diesen beiden Textemen.[28]

---

[24] (VAB75)

[25] in der aktuellen Referenzimplementation ist dies (noch) nicht der Fall

[26] Auf die ersten beiden Texteme des Beispieltextes (Anhang B) angewandt wären dies 1 Relation (Begriff – Begriffe) bei insgesamt 4 Stützwörtern, K_s in diesem Fall 1/4. Für die vollständige Tabelle siehe Anhang B.2, S. 82

[27] Vgl. dazu den Verlauf der beiden Koeffizientenwerte in Abbildung B.1, S. 84

[28] Tabelle B.3, S. 83 bietet eine Übersicht über die lokale Verflechtung aller Sätze untereinander. Der für jeden Satz spezifische mittlere Verflechtungsgrad ist in der Tabelle B.2, S. 82 aufgeführt.

Der lokale Koeffizient führt, falls erweitert, zu einem aufschlußreichen Kennzeichen der Texteme, indem man über das arithmetische Mittel einen Parameter der jeweiligen Verflechtung eines Satzes mit allen anderen Sätzen des Textes erhält:[29]

sei K_l für ein beliebiges Textem T_i aus der Menge aller Texteme T_r:

$$K_l(T_i) = \frac{R(T_i, T_j)}{N(T_i, T_j)}$$

mit $i \neq j$, dann berechnet sich der Grad der semantischen Verflechtung V des Textems T_i aus dem arithmetischen Mittel:

$$V = \frac{\frac{K_l(T_i,T_{j1})}{N(T_i,T_{j1})} + \frac{K_l(T_i,T_{j2})}{N(T_i,T_{j2})} + \cdots + \frac{K_l(T_i,T_r)}{N(T_i,T_r)}}{M-1} =$$

$$= \frac{1}{M-1} \cdot \sum_{j=1}^{M-1} \frac{R(T_i, T_j)}{N(T_i, T_j)}$$

Ein exemplarischer Vergleich[30] der semantischen Belastung und des lokalen Koeffizienten am Beispieltext[31] zeigt, dass einige Texteme eine deutliche Relevanz für das Thema des Textes aufweisen. Stellt man diese Ergebnisse einer Auswahl von Textemen gemäß dem Satz nach AGRICOLA gegenüber, demzufolge diejenigen Sätze eine Art "'Brückenkopf"-Funktion übernehmen, in welchen sich die Haupttopikketten (in diesem Fall $T_{k1}, T_{k3}, T_{k5} und T_{k6}$) kreuzen, so bestätigt sich das Ergebnis ein weiteres Mal.

Wie der lokale, so kann auch der **regressive Koeffizient** herangezogen werden, um die Grenzen zwischen stärker zusammengehörenden Textemen aufzuzeigen. Sein Wert sinkt proportional zur Abnahme der Kohärenz zwischen Textemen:

$$K_r = \frac{K}{M-1}$$

Dabei ist K die Zahl aller Isotopiebeziehungen zwischen den Lexemen einer beliebig langen Textemfolge und M steht für die Anzahl der Texteme dieser Folge.

Eine entsprechende Anwendung des regressiven Koeffizienten deckt insbesondere die Abschnitte des Textes auf, die semantisch enger beieinanderliegen. Über ihn lassen

[29] Diese Erweiterung ist vom Verfasser eingeführt worden. Mit dem arithmetischen Mittel des lokalen Koeffizienten läßt sich ein noch deutlicheres Bild des thematischen Gewichts der Texteme nachzeichnen, vgl. das bereits angesprochene Diagramm in Anhang B, Abbildung B.1, S. 84

[30] Vgl. Tabelle B.1, S.82

[31] Siehe S. 79

sich Grenzen thematischer Absätze aufzeigen. Damit kann ein weiteres Problem des Textzusammenfassens auf der Basis einfacher Satzextraktionen vermieden werden: Zusammen mit den thematisch relevanten Sätzen werden zusätzlich diejenigen Textemfolgen aus einem Text extrahiert, die für das Verständnis des "Themasatzes" von entscheidender Bedeutung sind. Oft verhindern aus dem Zusammenhang gerissene selbst inhaltsrelevante Sätze eine akzeptable Textzusammenfassung[32].

Textemfolgen (Beispiele)	K_r
$T_{9,10}$	2
$T_{7,8,9}$	1,5
$T_{5,6,7,8}$	1,3
$T_{2,3,4}$	1,5
T_{gesamt}	2,0

Tabelle 4.2: Textemfolgen, mit entsprechendem regressiven Koeffizienten

Trotz einer wesentlich längeren Satzfolge ist der K_r zwischen Satz $5-8$ kleiner als der K_r für das Textempaar 9 und 10: Ein Vergleich zwischen Textem 7, 8 und 9 zeigt, dass der regressive Koeffizient langsam absinkt, bis er zwischen den Textemen 5, 6, 7 und 8 auf ein Minimum herabfällt und erst zwischen Satz 2, 3 und 4 wieder ansteigt. Auf diese Weise kann die Strukturierung des gesamten Textkörpers anschaulich beschrieben werden und bei der Extraktion von Themasätzen berücksichtigt werden, indem Themasätze zusammen mit ihrem Kontext ausgewählt werden.

Aus diesen Tabellen wird ersichtlich, dass einige Sätze den Gesamttext dominieren. Ihr lokaler Koeffizient sowie die semantische Belastung weisen darauf hin, dass sie für die Themenstruktur des Textes von entscheidender Bedeutung sind (T_2, T_9, T_{12}). Für eine Thema-Ermittlung via Extraktion relevanter Texteme sind diese Ergebnisse mehr als genügend und beweisen die Mächtigkeit des Isotopieansatzes in Bezug auf die Thema-Analyse. Über mehrere Abhängigkeitsrelationen zwischen Textem und Topikkette kann das thematische Gewicht eines Textems präzise angegeben werden. Dabei stellt sich heraus, dass nicht unbedingt die Sätze mit den meisten Stützwörtern dominieren. Spielen die darauf aufbauenden Topikketten keine Rolle (d.h., sind sie zu kurz, um sich entscheidend für das Thema zu erweisen), so rückt die Quantität der Textaktanten das Textem nicht zwangsläufig in den Vordergrund.

Über die Verteilung und Struktur der Topikketten lassen sich nicht nur die wichtigsten Sätze eines Textes bestimmen. In einem nächsten Schritt, jenseits einer reinen Satzextraktion zur Kondensation des Textinhaltes, könnten Topikkettenrelationen sogar die Grundlage für den Aufbau einer in natürlicher Sprache generierten

[32] Vgl. das Problem des "shallow coherence smoothing" (Man01, S. 78)

Zusammenfassung des Textes bilden. Ein viel schwerwiegenderes Problem aber als das Nachzeichnen der thematischen Achsen im Text besteht jedoch darin, die Topikketten überhaupt aufzustellen. Wie ein solcher Vorgang, der die Semrekurrenz-Beziehungen automatisiert erfassen und die entsprechenden Lexeme in Ketten aufteilen muß, auszusehen hat, soll im Folgenden näher erörtert werden.

4.3 Graphische Darstellungsmöglichkeiten

Da Agricola den Informationskern als semantische Grundstruktur des Textes konzipiert hat, handelt es sich bei ihm um kein simples Textkondensat, kein reines Textthema, welches aufzufinden allerdings Ziel der vorliegenden Arbeit ist. Der Informationskern weist nach Agricola deshalb eigentlich genügend Merkmale auf, um aus ihm – inhaltlich identische – Paralleltexte konstruieren zu können. Mithin muß der Informationskern auch Strukturkomponenten sowie minimale Relationen beinhalten, die die inhaltliche Grundstruktur des Textes bereits vorwegnehmen.

Trotz dieses weiten Geltungsanspruches der Informationskerntheorie liefert sie (wie zu zeigen sein wird) auf Textkompressionsverfahren und Textthema-Analysen beschränkt, qualitative Resultate; selbst wenn die Möglichkeit anschließender Textgenerierung in dieser Arbeit ausgeklammert wurde und die Konnektorenstruktur zwischen den Sätzen[33], die für den Wiederaufbau eines Textes unerläßliche Informationen beinhalten, ignoriert wurden.

Der semantische Zusammenhang zwischen den Textaktanten findet in Agricolas Darstellung eine anschaulich graphische Visualasierung.[34]Diese bildhafte Übersicht über das "Gewebe" der im Text verflochtenen Topikketten dient lediglich der Verdeutlichung des bisher Besprochenen. Über eine Projektion der einzelnen Satzrelationen zwischen den Stützwörtern der Topikketten im Satz, schält sich das Gefüge des Informationskerns heraus. Bei den semantischen Dominanten des Beispieltextes[35] handelt es sich um:

A = Wörter
B = Begriffe
C = Netz

[33] Die RST Analyse beruht im wesentlichen auf Konnektorenstrukturen. Der Nutzen, der aus einer Verbindung von Isotopieketten und Miteinbeziehung von Konnektoren entspringen könnte, wurde von Barzilay und Elhadad angesprochen und in (IM99, S. 112) angedeutet. Auch vor dem Hintergrund dieser vielversprechenden jüngeren Forschungsbemühungen wäre eine eingehendere Untersuchung des Agricolaschen Konzepts vonnöten.

[34] Siehe Abbildung 4.3, S. 37

[35] Siehe Anhang B.1, S. 81

D = Wissenschaftler
E = Thesaurus
F = Sprache
G = Programm
H = Bedeutung

Die Vernetzung und Ordnung des Informationskerns (entspricht den syntaktisch-semantischen Relation der Topikketten zueinander) kann somit an der Graphenstruktur "abgelesen" werden. Baut man darüber hinaus auch die "logisch-relationell-kompositorischen"[36] Bestandteile, d.h. die Konnektoren, die die Texteme miteinander logisch verknüpfen, in diesem Schema ein, so gelangt man nach AGRICOLAzu einem "Konzentrat des Mitteilungsgehaltes eines Textes".

Ein Auszug aus dieser Konnektorenstruktur und ihre Berücksichtigung am vorliegenden Text wäre in dem Verhältnis angedeutet, auf dass durch K_1 hingewiesen wird. Konnektor K_1 stellt zwischen Textem T_1 und T_2 eine paraphrastische Beziehung her:

$K_1 \rightarrow PAR(T_1, T_2)$

während die logische Beziehung zwischen Textem T_1 und T_3 mit einer "Satzverknüpfungsrelation"[37] der Art

$K_1 \rightarrow MITTELS/DURCH(T_1, T_3)$

beschrieben werden könnte.

Wie erwähnt, sah AGRICOLA als Ziel dieses formalen Beschreibungsmodells (das sich zur graphischen Veranschaulichung semantischer Kerntrukturen im Text anbietet), die Grundlage für eine natürlich sprachliche Textgenerierung. Auf die Vor- und Nachteile dieses Modells bei der "Wiederentfaltung" des Textthemas zu einem Textganzen kann im Rahmen dieser Arbeit nicht weiter eingegangen werden. Die folgende Grafik veranschaulicht jedoch das Resümierte und schafft einen Überblick über die Informationskernstruktur des Beispieltextes[38].

4.4 Überlegungen zu einer semantischen Matrix

Nachdem deutlich geworden ist, wie die tragenden Textsegmente zur Textthemagewinnung beisteuern, stellt sich die Frage, unter welchen Bedingungen ein automatisiertes Aufstellen von Topikketten zu bewerkstelligen ist.

[36] (Agr79, S. 55)
[37] ibid.
[38] Siehe Abbildung S. 37

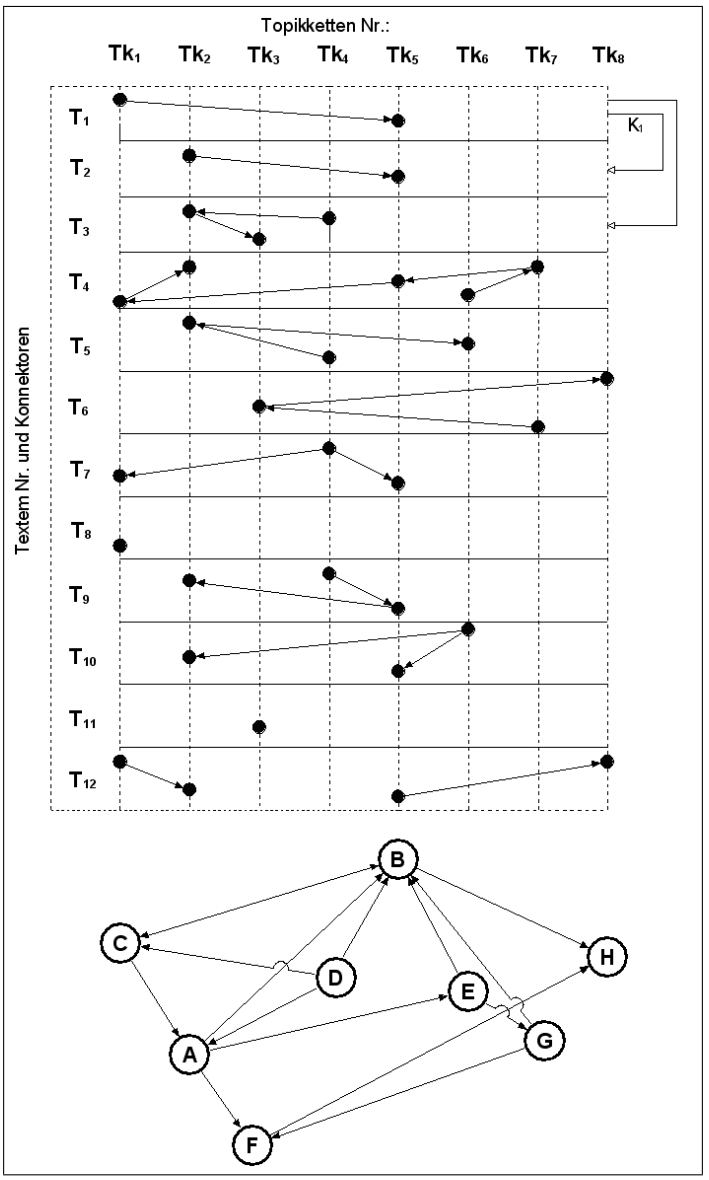


Abbildung 4.3: Vereinfachte graphische Darstellung des Informationskerngefüges

Dabei müssen die vielfältigen semantischen Beziehungen zwischen allen Lexemen des Textes[39] berücksichtigt werden. Eben hier läßt sich ein rein objektives Vorgehen nur schwer erzielen, scheint das Fatum der Semrekurrenz in weiten Teilen vom Vorwissen eines kompetenten Sprachteilnehmers abzuhängen, der allein entscheiden kann, ob zwei Wörter im einen Fall korrellieren, im anderen Falle aber distinktiven Topikketten angehören.

Nun hat die germanistische Linguistik bereits mehrere Versuche unternommen, die semantischen Beziehungen zwischen den Lexemen einer Sprache aufzuzeichnen und

[39] Wie bereits deutlich wurde, reichen generell Ketten aus Substantiven und Verben aus, um relevante Ergebnisse zu erzielen (IM99, S.113)

an diese vielfältigen Beziehungen objektive Maßstäbe anzulegen[40]. Die Lexikographie hat zu diesem Zweck Begriffswörterbücher und Thesauri hervorgebracht, die die synonymischen, antonymischen, hyperonymischen etc. Beziehungen des deutschen Wortschatzes katalogisieren und − da sie in erster Linie als rhetorische Hilfsmittel konzipiert sind − hauptsächlich synonyme Lexeme alphabetisch sortieren bzw. nach Oberbegriffen hyperonymisch ordnen.

Im Fall der Begriffswörterbücher wurden bei der semantischen Klassifikationen des deutschen Lexeminventars je nach Betrachtungsweise entweder logische, naturwissenschaftliche oder philosophische Ordnungsprinzipien eingesetzt. Begriffswörterbücher, die den Wortschatz einer Sprache nach Kategorien differenzieren, könnten also herangezogen werden, jedes Lexem darauf zu prüfen, ob weitere Lexeme derselben Subkategorie im Text vorkommen, um diese dann zu einer gemeinsamen Topikkette zu verknüpfen.

Mit Recht weist Kühn jedoch darauf hin, dass solche Klassifizierungen der Sprache nach sprachexternen Kriterien nur Projektionen einer bestimmten Metasemantik (z.B. "die Welt aus naturwissenschaftlicher Perspektive") sind:

> "Während alle bisherigen Begriffsschemata ... der Sprache *von außen* ein fremdes, statisches Ordnungsgefüge aufzudrängen, ergeben sich die Begriffskategorein und Begriffsrelationen [...] aus der kommunikativen Funktion der Sprache selbst."[41]

Diese funktionale Abhängigkeit, so Kühn hätte den unschätzbaren Vorteil, dass das Einteilungssystem nicht statisch wäre, sondern über eine Rückkoppelung Veränderungen im lexikalischen System berücksichtigt werden könnten. Eine solche Organisation des Lexeminventars muss als anzustrebender Idealfall behandelt werden, der dem kognitiven, flexibel organisierten Wortschatz bzw. Begriffsnetz sehr nahe kommt und der – durch alltägliche Erfahrungen – einer steten Neuordnung unterworfen ist.

Zu den bisherigen lexikographischen Versuchen, allgemeingültige (statische) Begriffsstrukturen aufzubauen, zählen :

alphabetisch: Der erste deutsche Thesaurus von Eberhard, Maasz und Gruber[42] gilt als Grundlage einer alphabetisch organisierten Synonymik.

[40] (**?**, S.68-72)
[41] (Küh79, S. 116)
[42] (JAE30)

philosophisch-phänomenologisch: ROGET[43], SANDERS[44] sowie später SCHLESSING[45] waren die ersten, die eine Einteilung der Sprache auf semantischer Ebene über eine allgemeine Begriffshierarchie anstrebten. ROGET entwickelte 6 Hauptklassen (Abstract Relations, Space, Matter, Intellect, Volition, Affections), die er in genau 1000 weitere Unterklassen teilte. SANDERS orientierte sich im Wesentlichen an ROGETS Einteilung, kam allerdings mit 688 Begriffen unter 7 Oberklassen aus (Abstrakte Beziehungen, Raum, Stoff, Erkenntnisvermögen, Wille, Besitz; Eigenthum, Empfindungen; Gefühle; Gemüthsbewegungen). SCHLESSINGS Begriffswörterbuch hingegen ist eine 1:1-Kopie des Thesaurus nach ROGET.

naturwissenschaftlich: Eine eigenständigere Leistung hat DORNSEIFF hervorgebracht, der unter der Aufstellung eines Begriffsystems stets die "Anwendung der Phänomenologie und philosophischen Ontologie auf die Sprache" versteht. Sein Begriffswörterbuch enthält 20 Hauptgruppen, die wiederum in $20-90$ Unterpunkte zerfallen können und die Natur von "allgemeinen Seinsbeziehungen" zum "Subjektiven" nachzeichnen. Im Vergleich mit rein philosophischen Klassifikationen sind seine Hauptgruppen bereits deutlich von naturwissenschaftlichen Kriterien geprägt (Anorganische Welt, Stoffe; Pflanzen, Tier, Mensch (körperlich); Raum, Lage, Form; ... Sichtbarkeit, Licht, Farbe, Schall, Temperatur, Gewicht ... Wirtschaft, Recht, Ethik usw.)

Alle Versuche, die Organisation interner Begriffsnetze auf ein intersubjektives Niveau zu heben, können also nur Kompromisse bleiben, vor allem dann, wenn man von einem einzigen Ordnungsprinzip ausgeht. Während ein Sprachteilnehmer das Erwähnen chinesischer Schriftzeichen mit dem Bereich der Gastronomie assoziiert (*China – Restaurant*) und in eine – für ihn – logische Relation bringt, wird ein anderer dieselben Zeichen (*China – Peking*) mit Reiseberichten und persönlichen

[43] MORRIS und HIRST konnten mit diesem Thesaurus Topikketten erzielen, die zu 90% mit den intuitiven Topikketten eines Sprachteilnehmers übereinstimmten. Siehe dazu (IM99, S. 113); eine Online-Version dieses Begriffswörterbuchs ist unter http://www.ai.mit.edu/people/wessler/thes.html zugänglich (26.05.2003)

[44] Sanders, D.: Deutscher Sprachschatz geordnet nach Begriffen zur leichteren Auffindung und Auswahl des passenden Ausdrucks. Ein stilistisches Hülfsbuch für jeden Deutsch Schreibenden. 2. Bde. Hamburg, 1873-7

[45] Schlessing, A.: Deutscher Wortschatz oder Der passende Ausdruck. Praktisches Hilfs- und Nachschlagebuch in allen Verlegenheiten der schriftlichen und mündlichen Darstellung. Für Gebiete aller Stände und Ausländer, welche einer korrekten Wiedergabe ihrer Gedanken in deutscher Sprache sich befleißigen. Mit einem den Gebrauch ungemein erleichternden Hilfswörterbuch. Eßlingen, 1881-1907

Erinnerungen in Verbindung bringen. Dem Bestreben relativ eigenständige objektive Oberkategorien zu finden und systematisch den Wortschatz einer Sprache danach hierarchisch durchzustrukturieren, haftet dabei stets der Makel einer gewissen Willkür an.[46]

Die unterschiedliche Strukturierung und Gewichtung interner Wortfelder trägt damit zu dem bekannten Problem der hermeneutischen Differenz bei. Es erklärt, warum der Rezipient nur einen Teil der vom Autor intendierten Bedeutung aus dem Bedeutungspotential des Textes "herauszulesen" vermag: In Einzelfällen wird seine Assoziation der Lexemzusammenhänge anders verlaufen – es entsteht eine abweichende Gewichtung von Textsegmenten und daraus resultierend werden andere Textstellen für wesentlich wichtiger oder interessanter aufgefasst. Deshalb herrscht bei kürzeren Texten oft Übereinstimmung zwischen den Rezipienten (z.B. bei einer Zeitungsmeldung) in Bezug auf die Bedeutung des Gesagten. Eine längere Abhandlung oder ein literarischer Text, der neben einigen sehr dominanten Topikketten noch von einer Unzahl kleinerer determiniert und beeinflusst wird, führt aber zu divergierenden Ansichten.

Da mit der Mächtigkeit des Isotopiemodells durchaus ein domänenunabhängiges Verfahren vorliegt, muss eine Möglichkeit gefunden werden, das dynamische Begriffssystem des menschlichen Lexeminventars auf eine semantische Datenbank abzubilden, die über einen Rückkoppelungsmechanismus verfügt und "lernfähig" ist und sich den Bedürfnissen des Nutzers anpassen können sollte. Eingegebene Beispieltexte müßten dann über die Topikketten-Analyse Themasätze auswählen und ein versierter Nutzer müsste die Datenbank mit seiner subjektiven Einschätzung der Ergebnisse so trainieren, dass sich das Begriffsnetz der Datenbank der neuen Domäne anpaßt. Dabei kann davon ausgegangen werden, dass dies insbesondere im Bereich von Fachtexten nötig sein wird. Die semantischen Relationen zwischen Vokabeln des alltäglichen Lebens (Auto, Fahrzeug, Transportmittel etc.) werden seltener Veränderungen ausgesetzt sein. Daher sollte die Datenbank auch ohne Training genaue Zusammenfassungen aus Texten liefern, die sich vor allem aus Wörtern des Grundwortschatzes zusammensetzen und wenig Fachvokabular aufweisen.

Elektronische Thesauri, auf die bei der angestrebten Topikkettenbildung zurückgegriffen werden könnte, sind im deutschsprachigen Raum selten. Zwei bestehende Lösungen, die im Folgenden überblicksartig vorgestellt werden sollen, decken sich leider nur unzureichend mit den Forderungen, die ein Topikkettenansatz an eine semantische Sprachdatenbank richtet.

[46] (Küh79, S. 117)

In Amerika wurde bereits früh die Bedeutung semantischer Wörterbücher (in elektronischer Form) für den Bereich des *information retrieval*, *data mining* und für computerlinguistische Forschungen erkannt: Am Cognitive Science Laboratory der Universität Princeton entwickelte man unter Leitung von Professor George A. Miller WordNet[47], ein lexikalisch-semantisches Netzwerk, das die Wortarten Substantiv, Verb, Adjektiv und Adverb in so genannten *Synsets* nach bestimmten Relationen organisiert.[48] Bei dieser Arbeit stand vor allem der pragmatische Aspekt im Vordergrund, einen möglichst umfassenden elektronischen Thesaurus für die englische Sprache zu entwickeln. Grundlage für die Arbeit an WordNet war neben sprachwissenschaftlichen Voraussetzungen auch eine psychologische: WordNet basiert auf Untersuchungen, wonach das Gedächtnis des Menschen in vergleichbaren Strukturen organisiert sein soll.[49]

Eine Beispielanfrage nach dem englischen Wort *car* beschreibt den Aufbau der Synsets:

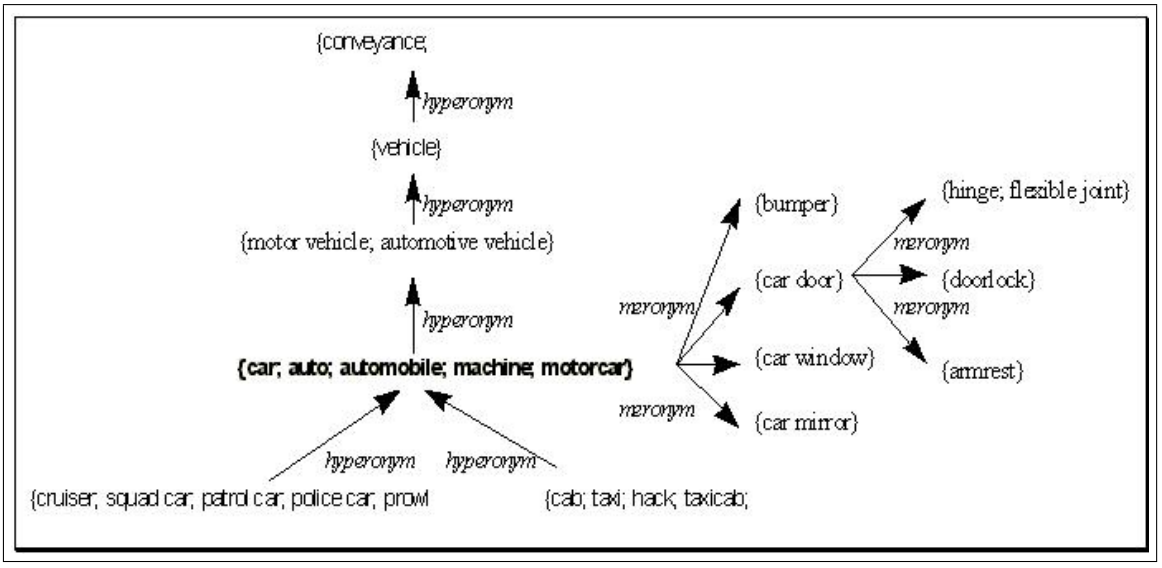


Abbildung 4.4: Synsets (hier für *car*) in WordNet

Inzwischen findet WordNet in zahlreichen Anwendungen der Computerlinguistik seinen Einsatz.[50] Eine europäische Initiative entwickelte anknüpfend an die englische Vorlage EuroWordNet[51]. Einige der wichtigsten europäischen Sprachen sind mitt-

[47] Siehe (Web00, S. 82) sowie http://www.cogsci.princeton.edu/ wn/index.shtml, (27.05.2003)
[48] (YAW96, S. 126)
[49] ibid.
[50] So basieren z.B.. die Arbeiten zur *lexical chain*-Analyse von Barzilay und Elhadad, Stairmand sowie Hirst und St-Onge (sic!) auf Wordnet
[51] Vossen, Piet: "Eurowordnet, final report", Technical Report D041, LE24003, LE4-8328, European Union, 1999

lerweile im EuroWordNet aufgenommen. Der europäische Thesaurus kann dabei auch über die Einzelsprache hinaus Beziehungen zu anderen Sprachen offenlegen. Grundlage für die deutsche Fassung von EuroWordNet war das am Institut für Sprachwissenschaften der Universität Tübingen entwickelte GermaNet[52]. GermaNet ist nachwievor ein alphabetisch geordnetes elektronisches Synonymwörterbuch, das angelehnt an das englische Orginal die Lexemrelatioen in Synsets organisiert. Bei der Entwicklung von GermaNet wurde auf folgende Thesauri zurückgegriffen:

- Deutscher Wortschatz Wehrle/Eggers
- Duden 8: Die sinn- und sachverwandten Wörter, und Duden 3: Bildwörterbuch
- Einsprachige deutsche Wörterbücher
- subjektives Sprachwissen der Projektmitarbeiter

Dabei unterstützen die Synsets in GermaNet Synonymie, Antonymie, Hyperonymie, Meronymie[53], Kausation, Derivation[54], Subereignis, Implikation sowie reguläre Polysemie.[55]

GermaNet wird bereits in einer zweiten Version als vollständig unabhängige Datenbank angeboten und wurde für diese Zwecke erweitert. Dem aktuellen Stand zufolge umfaßt der elektronische Thesaurus 41777 Synsets (davon allein 27824 Substantive) mit insgesamt 60646 Synonyma.

Beim Aufbau der Datenbank wurde der Wortschatz auf semantische Feldern aufgeteilt. Diese Ordnungsinstanzen, die bei der Entwicklung von GermaNet zum Tragen kamen, entsprechen zum Teil altbewährten Klassifikationen der Begriffswörterbücher. Zu den semantischen Feldern (den sogenannten "Tops") zählen: Artefakt, Attribut, Besitz, Relation, Geschehen, Form, Gefühl, Gruppe, Körper, Kognition, Kommunikation, Menge, Mensch, Motiv, Nahrung, natürlicher Gegenstand/Phänomen, Ort, Pflanze, Substanz, Tier, Zeit.

Die Gewichtung der einzelnen Relationen zwischen den Lexemen ist auch beim Menschen einer ständigen Veränderung unterworfen. Dabei entscheiden die Relationsschemata (Synonymie, Antonymie etc.) nur über die Möglichkeit der Verknüpfung – von der internen Bewertung dieser Beziehungen hängt jedoch das gesamte Begriffslexikon ab. So hätte WordNet im Textbeispiel[56] die Begriffe *Wort* und *Sprache*

[52] (BH01)

[53] Teil-von-Beziehung

[54] semantisch

[55] erlaubt Generalisierung und Prototypenbestimmung. Eine umfassende Klassifikation von Relationen zwischen lexikalischen Einheiten bietet Kühn (Küh79, S. 124-134)

[56] Siehe 79

nicht getrennt, sondern als engverwandte Einheiten zu einer Kette verbunden. Die Distinktivität dieser Begriffsphären hängt eben von der Gewichtung solcher semantischen Relationen zwischen den Lexemen ab und bedarf eigentlich – zwingend – einer dynamischen Repräsentation. WordNet liefert jedoch keine Möglichkeiten für die Gewichtung der Relationen; auch spielen dynamische Aspekte im WordNet-Thesaurus keine Rolle.

Ein weiteres deutschsprachiges Projekt, bei dem eine elektronische semantische Datenbank entwickelt wurde ist *SemDB*, eine morphosyntaktische Datenbank die semantische Klassen beinhaltet und im Rahmen des Verbmobil Dialogübersetzungssystems[57] entstand. Dieser Datenbank kam im Rahmen des modularen Ansatzes von Verbmobil die Aufgabe zu, semantischen Informationen, die aus der Satzanalyse des Eingangssatzes gewonnen wurden, so abzulegen, dass sie beim Generieren eines Antwortsatzes in der Zielsprache wieder herangezogen werden konnten.[58]

Zu diesem Zweck wurde eine Reihe von semantischen Klassen aufgestellt, die im Wesentlichen aus morphosyntaktischen Kriterien aufgebaut wurden und vererbungsbasiert verschiedene syntaktische Merkmale eines Lexems markieren. Eine Klassendefinition in SemDB weist die folgende Struktur auf:

```
class intransitive_c :< verb_c >: %Intransitives Verb
        semclass: intransitive_verb &     %Semantische Klasse
        predscheme: 'L,I' &               %Schema Prädikatname
        predscheme_a1: 'L, I, I1' &       %Schema für das Argument
        role_a1: (arg1 \ arg2 \ arg3).    %Themat. Rolle der Argumente
```

Ein Datenbankeintrag für das Verb "kommen" beeinhaltet so z.B. eine semantische Kurzklassifikation der Art:

```
sem_lex(Cat, kommen) short_for
        intrans_verb_sem(Cat, kommen, (space_time), [arg1]).
```

Die semantischen "Sorten" (semantic sort) sorgen dabei für eine gewisse Grundklassifikation der rund 7800 in SemDB enthaltenen Wörter und würden insbesondere für eine Topikkettenanalyse eine Rolle spielen. Andere semantische Begriffskategorien

[57] Das Verbmobil-System erkennt gesprochene Spontansprache, analysiert die Eingabe und erzeugt einen Satz in der Zielsprache, der dann über ein TTS Modul ausgegeben wird: Siehe URL (http://verbmobil.dfki.de/)

[58] (KW01, S. 3)

enthalten z.B..: Symbol, geo_location, Location, Field, info_content, communication situation etc.

SemDB fußt mit seinen 30 semantischen Feldern[59] auf einer wesentlich rigideren Ontologie (im Vergleich zu ROGETS 1000 Begriffsklassen), während GermaNet eine größere Anzahl lexikalischer Relationen berücksichtigt.

Für die Erstellung von Topikketten muß ein Zwischenmaß an Flexibilität und Stabilität erreicht werden: Um bei einer großen Anzahl unterschiedlicher *domains* und Textsorten qualitativ gleichwertige Ergebnisse zu erzielen, sollte die Verknüpfung von Topikketten selbst bei sehr weiten, assoziativen Beziehungen zwischen Lexemen funktionieren; allein deshalb schon, weil es gilt, dass fehlende Denotatswissen wettzumachen. Untersuchungen zur Organisation der Struktur des mentalen Lexikons unterstützen diese Auffassung enger assoziativer Verbindungen:

> "From the standpoint of retrieval of information in an associative memory, the small-world property of the network represents a maximization of efficiency: on the one hand, similar pieces of information are stored together, due to the high clustering, which makes searching by association possible; on the other hand, even very different pieces of information are never separated by more than a few links, or associations, which guarantees a fast search."[60]

Dieser Vorteil kann sich leicht in einen Nachteil verkehren, und zwar dann, wenn die semantische Nähe zwischen Stützwörtern des Textes zu Verbindungen führt, die für den Gegenstand des Textes völlig unangemessen sind. Ein Text über *Schwermetalle* und *Natur* z.B. behandelt diese Begriffssphären als einander gegenüberzustellende Topikketten. Ein Thesaurus wie GermaNet oder WordNet würde aber über die enge Relation zwischen Begriffen wie Metall und Natur diese zu einer gemeinsamen Kette zusammenfügen wollen.

Einen Ausweg aus diesem Dilemma, dass oft durch die polyseme Struktur der Wörter weiter verschärft wird, haben ELHADAD und BARZILAY in der Einführung einer heuristischen Bewertungsmethodik für die Gewichtung aller computablen Ketten vorgeschlagen: So berechnet ihr Programm alle denkbaren Kettenvarianten gleichzeitig und baut für jede Kette eine eigene Gewichtung nach der Art der Relation und

[59] (FX01)

[60] Zur *small-world* Eigenschaft des mentalen Lexikons und der Verbindung zwischen Assoziation und *high density clustering* vgl. (AEM02, S. 3) siehe auch http://www.wissenschaft-online.de/abo/ticker/595348, (14.05.2003) sowie (AB01,)

der textuellen Nähe von Stützwörtern auf. Ab einem Schwellenwert werden diejenigen Ketten ausgewählt, die die höchste Wertung erlangt haben, und das Programm verwirft den Rest, um die Rechenzeit nicht ausufern zu lassen.[61]

Dieser freie Assoziationsgrad bei der Topikkettenbildung, der also für eine enorme Domänunabhängigkeit sorgt, schafft Probleme bei der Aufstellung relevanter Isotopieverhältnisse.

Verknüpfungen, die zu spekulativ sind, könnten aber auch unmittelbar durch die Angabe einer Mindestlänge der Topikkette eleminiert werden. Dabei muß ein Verhältnis ermittelt werden, der die Mindestlänge der Topikketten von der Textlänge abhängig macht, so dass in kürzeren Texten Topikketten erst ab zwei semantisch nahestehenden Lexemen erzeugt werden, in längeren Texten hingegen Topikketten erst ab 5 oder mehr korrellierenden Lexemen aufgebaut werden. Wird der Schwellenwert geschickt gewählt, so sollte dieses Auslesekriterium dafür sorgen, dass nur die "stärksten", d.h. semantisch eindeutigen Ketten überleben. Würde ein Text also mehrmals das Lexem *Schwermetall* verwenden, so würde sich daraus eine eigene Kette bilden. Es könnte jedoch auch dazu führen, dass zwei sehr enge Ketten verschmolzen werden. Dies bereitet nur dann Probleme, wenn diese Ketten eigene thematische Achsen bilden.

Ein anderer wichtiger Faktor ist die Topikkettenstabilität (nur Lexeme mit expliziten semantischen Relationen werden in Topikketten eingefügt oder untereinander verbunden). Sie sorgt dafür, dass nicht wahllos Lexeme verknüpft werden, deren Zusammenhang zu assoziativ wäre. Auf existente Lösungen zurückgreifend, bietet GermaNet die umfassendsten Voraussetzungen für eine Topikkettenanalyse, wobei man jedoch beachten muß, dass ein großer Bestand an vernetzten Lexemen mit einer mangelnden Dynamisierung[62] der Wortfelder und ihrer Verknüpfungen (Rückkoppelung durch den Nutzer) erkauft wird.

[61] (IM99, S. 114)

[62] http://www.visualthesaurus.com/online/index.html

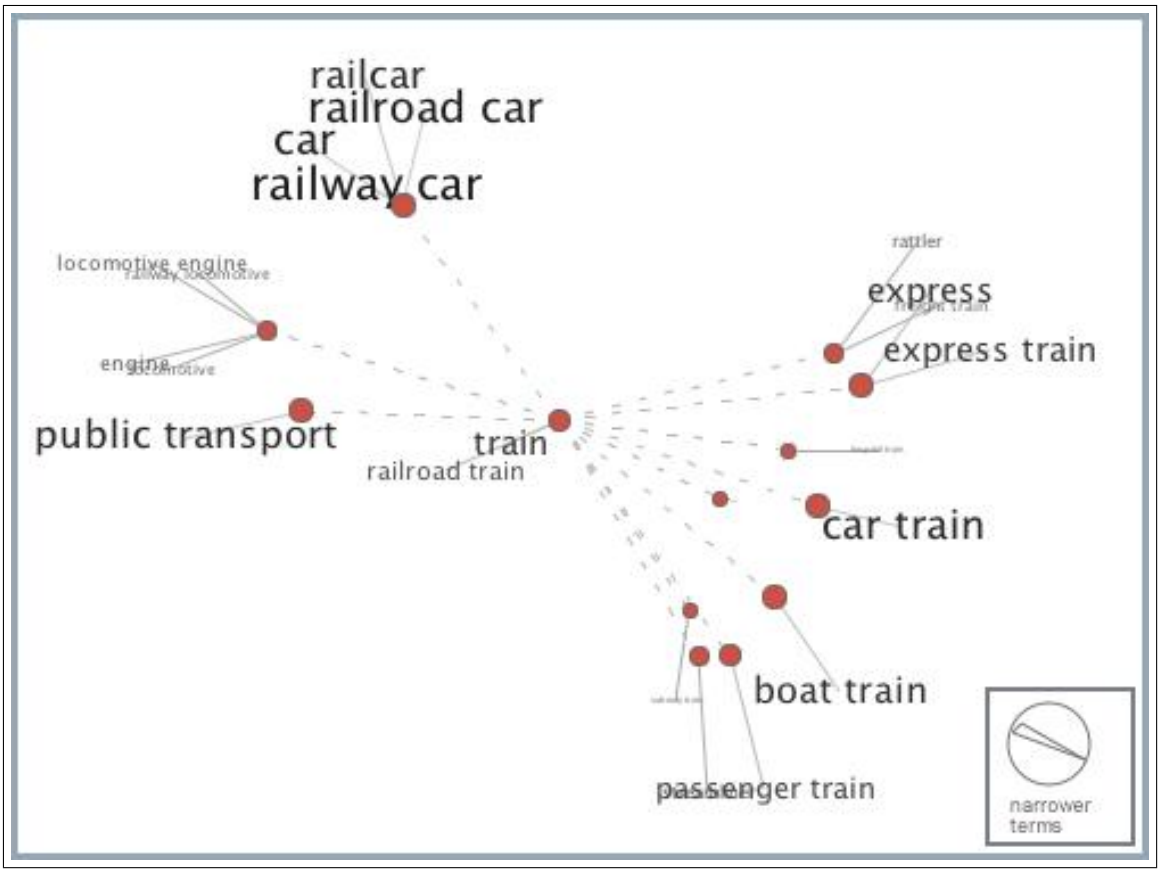


Abbildung 4.5: Visualisierte WordNet-Relationen am Substantiv *train*

5 Textthema-Analyse

5.1 Algorithmus

Aus den bisherigen Ergebnissen kann ein Algorithmus zur Analyse des Textthemas auf der Grundlage der Textisotopie abgeleitet werden. Ein prozedurales Vorgehen[1], das den Grundannahmen vorangehender Kapitel folgt, kann entweder dazu verwendet werden, einen Schlüsselsatz (bzw. eine Schlüsselsatzmenge) aus dem Text zu extrahieren oder eine Kurzfassung zu generieren. Zwecks Exemplifizierung dieses Ansatzes soll zunächst die einfachste Form des Algorithmus implementiert werden.[2]

1. Zuerst müssen alle Stützwörter aus den einzelnen Textemen ausgewählt werden. Diese Aufgabe kann von einem modifizierten Part-of-speech-Tagger (PoS) übernommen werden oder an ausgewählten Beispieltexten manuell erfolgen. Vorzugsweise kommen als Stützwörter Substantive und Verben in Betracht, da sie als Grundlage des Präzidierens für den Kern des Textems stehen und – im Verhältnis zu den anderen Wortarten – den größten Anteil an der Proposition eines Textems liefern. Eine vereinfachte Form dieses Algorithmus verzichtet auf die Aufnahme von Verben in Topikketten.[3]

2. Jedes Stützwort eines Textems muß einer Kette zugeordnet werden. Ist keine entsprechende Kette vorhanden, wird mit dem Stützwort eine neue Kette

[1] Für das entsprechende Flußdiagramm siehe 52

[2] Für die Generierung eines automatischen *abstract* wäre die Informationskerntheorie ein geeigneter Ausgangspunkt. Eine Untersuchung in dieser Hinsicht wurde selbst von Agricola seinerzeit angestrebt, aber bisher nicht unternommen.

[3] Um brauchbare Ergebnisse zu erzielen reichen Substantive vollkommen aus, vgl. (IM99, S. 115); zudem beschleunigt diese Beschränkung den Algorithmus

eröffnet. Ist die Anzahl der Kettenglieder nach Abschluß der Analyse des Gesamttextes kleiner als ein heuristisch festzulegender Schwellenwert, dann wird die Kette terminiert und in die weitere Textanalyse nicht einbezogen.

3. Im Kernanalyseprozess werden die Topikketten ihrer Länge entsprechend geordnet. Als Haupttopikketten gelten dabei jene Ketten, bei denen sich ein deutlicher Abstand zur Länge der restlichen Topikketten feststellen läßt.

4. Es werden diejenigen Sätze ermittelt, auf denen die zwei längsten Haupttopikketten ruhen. Aus dieser Textemmenge werden wiederum all diejenigen Sätze ausgewählt, durch die die drittlängste Topikkette verläuft. Dieser Prozeß setzt sich fort, bis der gefundene Satz von keiner weiteren Haupttopikkette mehr durchlaufen wird. Aufgrund der Valenz der Satzverben wird die Tiefe dieses Algorithmus selten über drei Topikkettenschnittmengen hinausgehen. Dies wird zum Problem, wenn der Text länger ist und sich aus vielen Topikketten zusammensetzt. Hier müßte der Gesamttext zuallererst in einzelne Absätze unterteilt werden, wobei die Absatzgrenzen – gemäß dem angeführten regressiven Koeffizienten – auch Sinnabschnitte repräsentieren. Diese Einzelabsätze werden dann in einem rekursiven Schritt wie Gesamttexte behandelt.

5. Bei dem gewählten Satz oder den gewählten Sätzen (im Falle einer Analyse mehrerer disjunkter Absätze) handelt es sich um den bzw. die gesuchten Schlüsselsätze.[4]

5.2 Pseudocode und Flussdiagramm

Eine Ermittlung der Schlüsseltexteme über den lokalen Koeffizienten – eine Bestimmung, die genauer wäre und sich in absoluten Zahlen ausdrücken ließe – führt wegen der polynomialen Komplexität eines solchen Algorithmus bei größeren Textmengen zu Leistungseinbußen:

$T(n) = O(n^2(k^2)) = O(n^2)$

wobei k eine Konstante ist, die sich aus dem Vergleich der (beschränkten) Anzahl von möglichen Stützwörtern ergibt und damit nicht weiter ins Gewicht fällt. Die quadratische Komplexität entsteht dadurch, dass ein Textem mit allen anderen verglichen werden muß. Der folgende Pseudocode beschreibt eine Textthema-Suche auf der Basis des lokalen Koeffizienten. Im Hauptteil des Programmes findet die eigentliche Berechnung statt. Dafür ist jedoch die Berechnung der gemeinsamen Stützwörter (R, entsprechend der Formel, S. 31ff.) nötig. Das geschieht im Unterprogramm

[4] Siehe Abb. Seite 52

satz_vergleich(). Dieses Unterprogramm liefert die Anzahl der Stützwörter zurück, die es über einen Zugriff auf die semantische Datenbank zuvor miteinander abgeglichen hat:[5]

```
procedure lokaler_koeffizient()
var:    i, j, n: integer;
        satz[n][n]: char;

begin
    for i := 1 to i <= n do
        for j := 1 to <= n do
            Rufe satz_vergleich() auf
            Dividiere die Anzahl der Stützwortrelationen zweier Texteme
            durch die Anzahl der ihrer Stützwörter insgesamt;
        end
    end
end

procedure satz_vergleich(i, j)
const: MAX
//maximale Anzahl der Stützwörter im Satz
var: k, l: integer;
begin
    for k := 1 to k <= MAX do
        for l := 1 to l <= MAX do
            Wenn eine Relation zwischen den
            Stützwörtern besteht then
                temp := temp + 1;
            end
        end
    end
    stützwörter_anzahl := temp;
    return(stützwörter_anzahl)
end

procedure Relation_besteht(satz eins, satz zwei)
begin
```

[5] Wobei n der Anzahl der Texteme entspricht.

```
    Öffne die SemantischeDB;
    Suche den ersten Satz;
    Suche den zweiten Satz;
        Wenn ein Eintrag von Satz eins mit Satz zwei
        übereinstimmt then
            return(true)
        end
end
```

Da sich jedoch die Ergebnisse aus der Schnittmenge der Haupttopikketten mit der Berechnung aus dem lokalen Koeffizienten decken[6], kann über die Schnittmenge der längsten Topikketten ein ähnliches Resultat bei einer günstigeren Komplexität von

$$T(n) = O(nk + n) = n$$

erzielt werden. Diese Obergrenze ergibt sich aus dem zusammengesetzten Hauptprogramm. Im ersten Schleifenkörper findet die Bildung der Topikketten linear statt, geht man davon aus, dass in einem Satz nur eine begrenzte Anzahl von Stützwörtern auftreten kann (also konstant ist) und die Topikkettenanzahl nicht wie in Barzilay und Elhadads Ansatz parallel berechnet wird, sondern über einem heuristischen Schwellenwert abgebrochen wird.[7] Nach der Bestimmung der zwei längsten Topikketten werden die Textthemasätze ausgewählt. Ohne Berücksichtigung eines Sortieralgorithmus und einer weiteren Feinregulierung der Themakondensation ist der Algorithmus von linearer Komplexität.

```
procedure SchnittHauptTopikketten()

const:  schwellenwert: integer;
var:    i, j, l, m, n, q :  integer;
        //l ist der Index der TK
        //m entspricht dem Index des Lexems in Textem n
begin
    for i := 1 to i <= n do
    //bis Textende n
        for j := 1 to <= m do
```

[6] Vgl. dazu S. 31

[7] Die Anzahl der Topikketten pro Absatz muß aufgrund der Stützwortanzahl und der Kohäsion des Textes endlich (klein) bleiben. Es bietet sich daher an, eine Textthemabestimmung auf Subtextgrößen vorzunehmen (die über den regressiven Koeffizienten zuvor ermittelt werden), um die Einzelergebnisse später zusammenzufassen.

```
        //jedes Stützwort m
            Topikketten()
            //Unterprogramm zur Topikkettenbildung
        end
    end
    for p := 1 to topikketten_anzahl do
    //Topikkettenauswahl
        Die Topikketten, die einen festgelegten Schwellenwert über-
        schreiten, werden der länge nach geordnet
        end
    end
    for k := 1 to n do
     Teste, in welchen Sätzen sich die 2 Haupt-TK treffen
     Wenn sie sich treffen:
       zielsatz_anzahl := zielsatz + 1;
    end
end
procedure LexemTopikkettenVergleich()
begin
    for l = 1 to topikketten_anzahl do
        if Besteht_Relation()
         Überprüfen ob Reiteration oder
         Semrekurrenz vorliegt. Wenn ja,
         return(tk);
        end
    end
end
```

Im zweiten Teil der vorliegenden Arbeit soll dieser zweite Algorithmus exemplarisch implementiert und auf einige ausgewählte Beispieltexte angewandt werden.

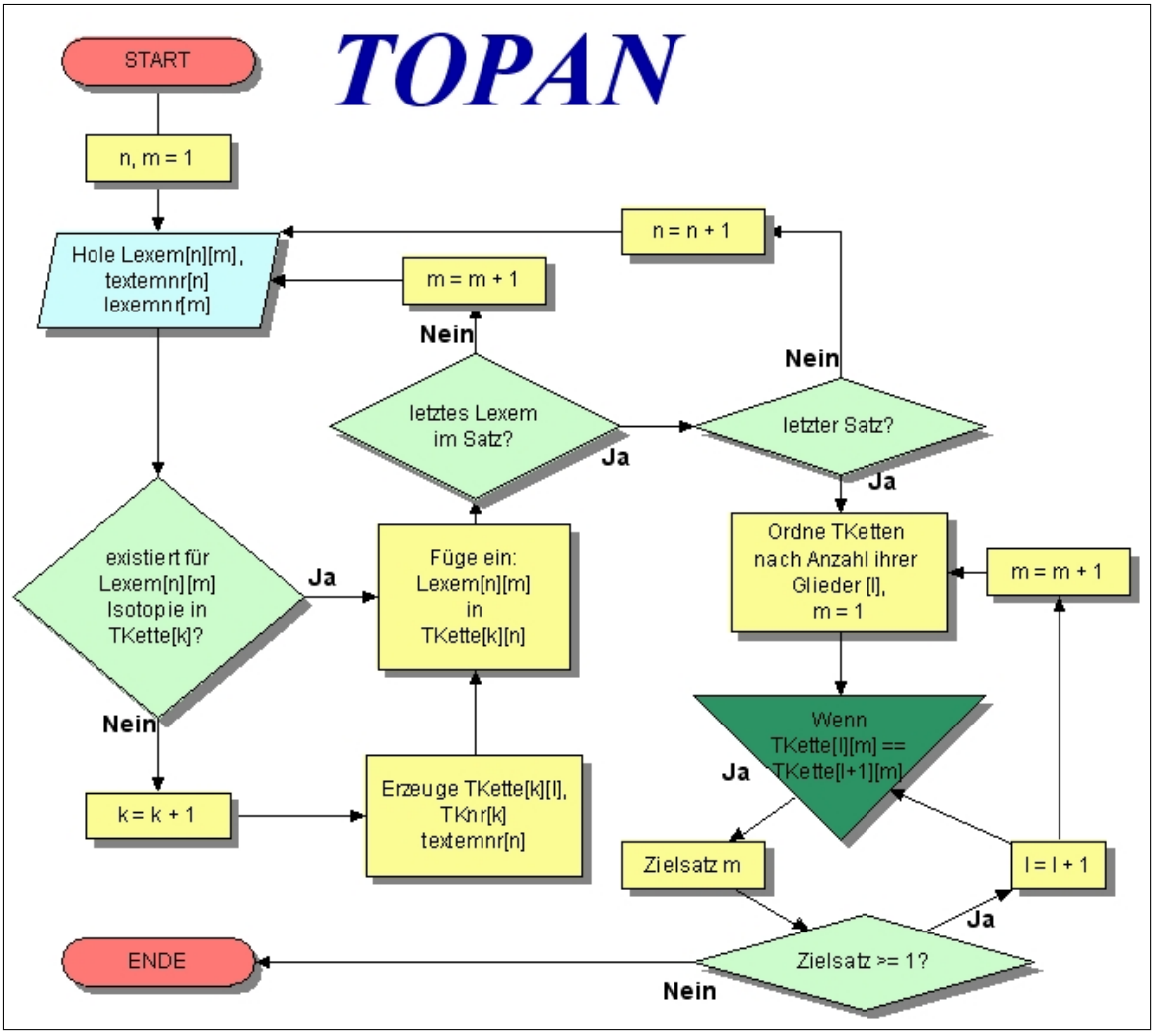


Abbildung 5.1: Funktionsweise des Algorithmus basierend auf den Schnittmengen der Haupttopikketten

Teil II

Implementation von TOPAN

6 Aufbau und Funktionsweise

6.1 Vorverarbeitung: PoS-Tagger

Das Topikketten Analyseprogramm "TOPAN" soll die im ersten Teil aufgeworfenen Hypothesen zum Informationskern – insbesondere dessen Anwendung auf die Extraktion der thematisch relevanten Schlüsselsätze – verifizieren.

Dazu muss ein beliebiges Textdokument im ersten Schritt einer "Normalisierung" unterzogen werden. Diese Normalisierung kann soweit gehen, dass allein die Grundstruktur eines jeden Satzes, bestehend aus Autosemantika und Verben, aus dem Satz "herausgeschält" wird.[1]

Für eine Demonstration der Funktionsweise der postulierten Textthema-Gewinnung mittels Topikkettenstrukturen – und um mehr als eine Demonstration kann es sich bei der folgenden Referenzimplementation allein aus Gründen des Zeitvolumens nicht handeln – ist es essenziell, zumindest eine Liste aller Substantive pro Textem/Satz vorliegen zu haben, da aus denselben die Topikketten gebildet werden müssen.

Diese Wortartanalyse, gemeinhin Bestandteil eines Textparsing-Prozesses, bei dem die einzelnen Lexeme ihrer Wortart und/oder Satzfunktion nach bestimmt werden, kann entweder über einen PoS erfolgen, oder manuell, d.h. per Hand vorgenommen werden.[2]

Für die deutsche Sprache bieten sich als Tagging-Programme u.a. "Brills Tagger", "Morphy" und "TreeTagger"[3] an.

[1] solch eine Normalisierung sieht Agricola für die Bestimmung seines Informationskerns vor (Agr79, S. 55)

[2] wobei manuelles Tagging höchstens bei Kleinsttexten sinnvoll ist.

[3] für Linux/Unix und Solaris Systeme

Brills Tagger[4] basiert auf einem regelbasierten Taggingalgorithmus und ist damit eine Alternative zu den verbreiteten statistischen Taggingprogrammen, die über verschiedene Näherungsverfahren und Kontextwahrscheinlichkeiten (z.B. Hidden Markov Model) die Wortart bestimmen. Dieser Tagger, dessen Sourcecode frei verfügbar ist und der sich damit für die Integration in eigene Softwareentwicklungen anbietet, hat den Nachteil, dass die OpenSource Variante an englischen Textkorpora trainiert wurde und man deshalb erst für das Deutsche aus manuell getaggten Texten einen Regelapparat aufbauen und verfeinern muss.[5]

Als Ersatz[6] für diesen genaueren regelbasierten PoS Tagger, soll als Vorbereitung für die Topikketten-Analyse an dieser Stelle das von Wolfgang Lezius[7] entwickelte Morphologie und Tagging Programm "Morphy" zur Anwendung kommen.[8] Morphy enthält ein Stammformenlexikon mit über 50.000 Einträgen (ca. 350.000 Vollformen), schlägt bei unbekannten Wörtern von selbst die wahrscheinlichste Wortklasse vor und unterstützt dank einer Eingabemaske eine flexible Lexikonpflege.

Mithilfe dieses PoS Taggers sollen ausgewählte Beispieltexten nach Substantiven ausgefiltert werden und normalisiert dem eigentlichen Programm übergeben werden. Dabei werden die von Morphy vorgeschlagenen Wortarten mitberücksichtigt.

Dieser erste Schritt in der Programmanwendung stellt sich dem Nutzer wie folgt dar[9]:

Über den Menüpunkt Datei|Öffnen wird der zu analysierende Text (eine txt-Datei) geöffnet. Diese wird im ersten Karteifenster sichtbar[10].

Für die weitere Nutzung des Programmes muss auf eine lemmatisierte Fassung zurückgegriffen werden, die in einem Vorverarbeitungsschritt mittels Morphy erzeugt

[4] Siehe http://www.cs.jhu.edu/ brill/home.html, (27.05.2003)

[5] Ein solches Vorhaben hat Megyesyi für das Ungarische umgesetzt: Siehe http://www.speech.kth.se/ bea sowie http://www.speech.kth.se/7Ebea/finalML.pdf, (20.05.2003)

[6] Die Computerlinguistik Abteilung der Universität Zürich hat den Brill-Tagger für das Deutsche trainiert. Der schweizer Tagger basiert auf einem Trainingskorpus von rund 58'000 Wörtern (Themenbereich: Jahresberichte der Universität Zürich). Als Tagset wurde das STTS gewählt, das rund 50 Tags umfaßt (plus Tags für Satzzeichen). Der Tagger kann jedoch leider nur online getestet werden, eine Eingabe ist auf 70 Wörter beschränkt. Siehe http://www.ifi.unizh.ch/CL/tagger/, (20.05.2003)

[7] http://www-psycho.uni-paderborn.de/lezius/, (20.05.2003)

[8] Eine Vorläuferversion von TOPAN beinhaltete einen vom Verfasser entwickelten rudimentären PoS. Für eine optimale Substantiverkennung mußte deshalb ein ausgefeilter PoS gefunden werden. Da es leider nicht möglich war C-basierte Open-Source-Tagger in die Applikation einzubinden, wurde in der aktuellen Programmfassung auf Morphy zurückgegriffen

[9] Siehe Abbildung 6.1, S. 85

[10] Siehe Abbildung 6.2, S. 57

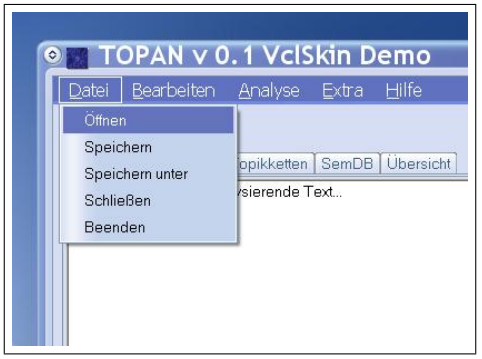


Abbildung 6.1: Öffnen des Zieltextes I

worden ist. Morphy kann ganze Dateien einer morphologischen Analyse unterziehen. Wählt man die Einstellung “Standard Ausgabe, Tabellenform”[11] so produziert das Taggingprogramm, neben anderen Ausgaben, auch eine Lemmata-Datei (*.lem), die allen im Text aufscheinenden Wortformen u.a. ihre entsprechenden Wortklassen zuordnet.

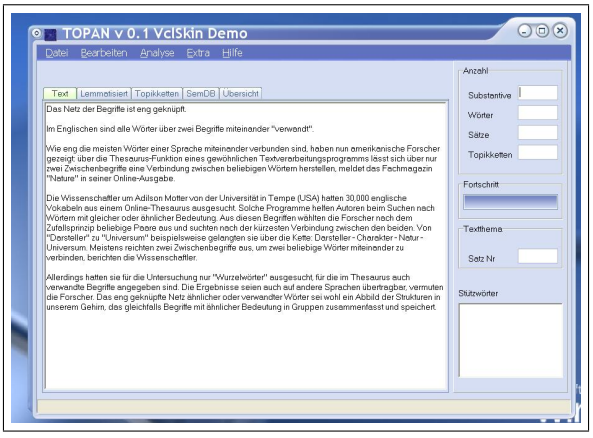


Abbildung 6.2: Öffnen des Zieltextes II

Diese Lemmataliste muss als nächstes in TOPAN geöffnet und weiterverarbeitet werden. Dazu wird mit dem Dialogfenster die entsprechende Lemmatadatei geladen[12] und erscheint im zweiten Karteifenster[13].

Die eigentliche Normalisierung, die für die Bildung der Topikketten erforderlich ist, greift auf die soeben geladene Lemmatadatei zurück. Über den Menüpunkt Analyse | Normalisieren werden aus der Lemmatadatei die Substantive satzweise ausgelesen und zeilenweise markiert wiedergegeben. Dabei ist hervorzuheben, dass TOPAN

[11] Siehe Abbildung 6.3, S. 58
[12] Siehe Abbildung 6.4, S. 58
[13] Siehe Abbildung 6.5, S. 59

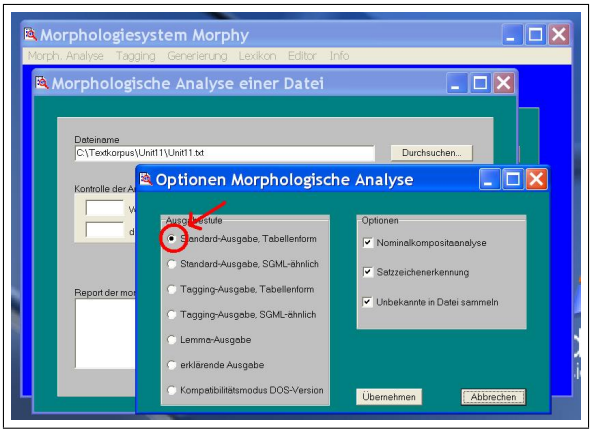


Abbildung 6.3: Erstellen einer Lemmata Datei

auch die von Morphy vorgeschlagenen Wortklassenzuweisungen untersucht und gegebenenfalls einbaut.[14]

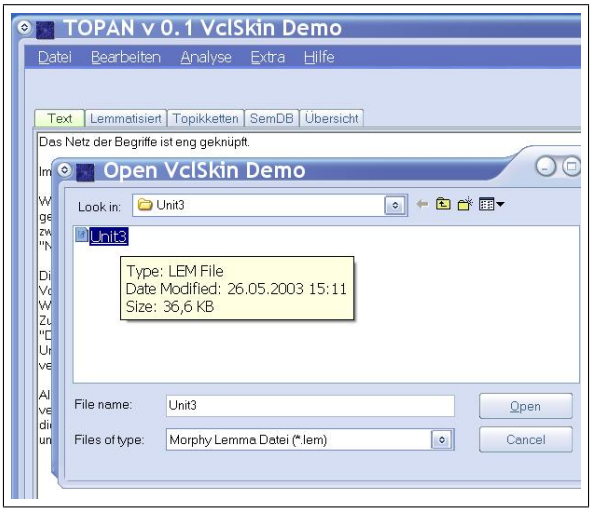


Abbildung 6.4: Laden der *.lem Datei

Zurückgreifend auf vorangegangene Überlegungen, denen zufolge eine Vorsegmentierung des Textes in Absätze von gleicher Themenstellung gerade bei umfangreicheren Texten Sinn machen würde, soll an dieser Stelle auf eine ähnliche Vorgehensweise bei Brunn, Chali und Pinchak hingewiesen werden.

Sie nutzen für diese Zwecke allerdings keine Topikketten, sondern einen *text segmenter* der von Choi vorgestellt worden ist.[15] Eine Weiterentwicklung von TOPAN müßte nach einer Untersuchung der Textsegmentierung entsprechend regressivem

[14] In der überwiegenden Zahl der Fälle trifft der von Morphy abgegebene Vorschlag zu. Bei unbekannten Worten handelt es sich zudem meist um fachspezifische Termini. Der normalisierte Text präsentiert sich abschließend wie in Abbildung 6.6, S.59

[15] (MB01, S. 1)

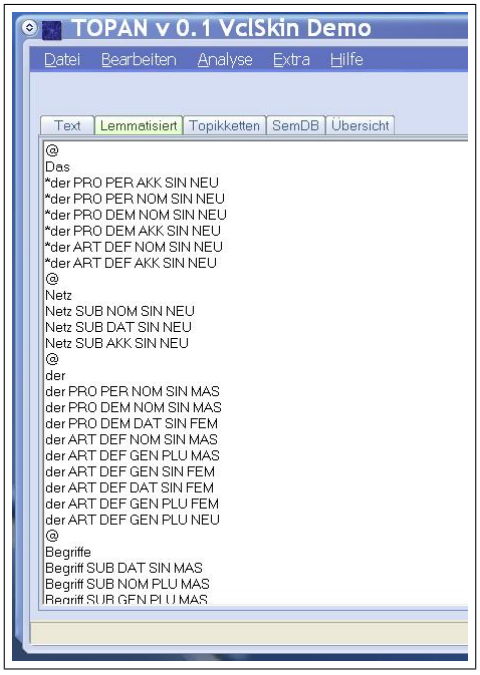


Abbildung 6.5: Lemataliste

Koeffizienten und CHOIs Algorithmus eine entsprechende Vorsegmentierung unterstützen.

Abbildung 6.6: Normalisierter Text

6.2 Semantische Datenbank

Wie der theoretische Teil gezeigt hat, steht und fällt die Themagewinnung vermöge Topikketten mit dem Umfang und der Dynamik einer semantischen Datenbank. Ihr Umfang garantiert Domänunabhängigkeit, ihre Flexibilität sorgt für die nötigen Veränderungen in der Gewichtung einzelner Relationen.

Aufgrund des Umstandes, dass es für die Deutsche Sprache kaum Begriffswörterbücher oder Thesauri in elektronischer Form gibt, erschwert sich insbesondere der Vergleich zwischen unterschiedlichen Organisationsstrukturen semantischer Netze. Eine solcher Vergleich wäre aber dringend erforderlich, soll die Stärke des hier vorgestellten Ansatzes in aller Breite evaluiert werden.

Grundlegend lassen sich zwei mögliche Herangehensweisen unterscheiden: Zum einen kann der Aufbau von Topikketten über explizite lexikalisch-semantische Relationen erfolgen. Dabei enthält die semantische Datenbank zu jedem Lexikoneintrag eine Liste mit möglichen Relationen (Hyponymie, Hyperonymie, Teil-Ganzes-Beziehungen etc.). Beispiele für eine derartige Sammlung lexikalischen Wissens liefern WordNet oder GermaNet.

Eine zweite grundsätzliche Organisationsform des semantischen Wissens könnte in einer Klassenhierarchie angeordnet sein. Dabei wird ähnlich wie in einem Begrifflexikon von übergeordneten Kategorien (Raum, Zeit, Mensch, Gefühl etc.) ausgehend eine Ausdifferenzierung des Wortschatzes vorgenommen. Bei jedem Lexikoneintrag wird nun in Folge geprüft inwieweit er der einen oder anderen Hierarchie näher steht. Als Folge entsteht eine binäre Liste, die ein Lexem jeweils charakterisiert:

Lexem	Raum	Mensch	Gefühl ⋯
	Ausdehnung ...	Geschlecht...	Trauer ⋯
Strecke	1	0	0 ⋯
Geist	0	0	1 ⋯

Tabelle 6.1: Semantische Datenbank in Form einer Klassenhierarchie

Beide Varianten der Wörterbuchstruktur haben ihre Vor- und Nachteile. Während Einträge synonymisch oder ähnlich relational verbundener Lexeme konkreter und Zusammenhänge fassbarer sind, ist der "Assoziationsraum" dieser Thesauri deutlich eingeschränkt. In jedem einzelnen Fall müssen alle möglichen Relationen erfaßt und eingetragen sein, damit Topikketten zwischen Wörtern gebildet werden. So wird z.B. das Wort *Typ* in keiner explizit lexikalischen Relation zu *Sprache* stehen. Dennoch ist ein Text vorstellbar, nach dem von einem *Sprachtypus* die Rede ist. Wie soll das Programm hier entscheiden, ob beide Worte in einer Isotopiebeziehung stehen?

In diesem Fall erweist sich das zweite Modell einer nach Begriffskategorien oder Klassen angeordneten semantischen Matrix als flexibler. Eine umfangreiche Anzahl an Klassen würde die einzelnen Lexeme voneinander unterscheidbar machen. Nur Synonyme oder begriffsverwandte Lexeme hätten in einer solchen Aufstellung eine vergleichbare "binäre ID", bestehend aus der jeweiligen Zugehörigkeit zur einen oder anderen Klasse. Ein zuvor festgelegter Prozentsatz der Übereinstimmung zwischen

zwei solcher Einträge entscheidet nun darüber, ob es zu einem *matching*, einem lexikalischen Kongruieren zweier Lexeme kommt. Ähnlich der Assoziationsweite eines menschlichen Rezipienten greift das Programm auf einen "vage" Nähe zwischen zwei Lexemen zurück und verbindet sie, wird ein Schwellenwert überschritten, zu einer gemeinsamen Kette.

Da dem Verfasser im Rahmen dieser Arbeit GermaNet leider nicht zugänglich war und ein hier undokumentiertes Vorläufermodell von TOPAN bereits auf eine semantische Datenbank nach Begriffkategorien getestet worden ist, soll für das folgende Programm eine Beispieldatenbank zum Einsatz kommen, die, in Anlehnung an GermaNet, jeden Lexemeintrag mit Wörtern aus den Relationen Überbegriff, Unterbegriff, Teil-Ganzes-Beziehung versieht. Als Quelle für diese Semantische Datenbank (SemDB)[16] wurden Einträge aus dem Duden der "sinn- und sachverwandten Wörter"[17] ausgewählt, sowie Ergebnisse des Online verfügbaren Leipziger Wortschatz-Projektes[18] mitberücksichtigt. Diese semantische Datenbank beinhaltet jedoch in der vorliegenden Form selbstredend nur Einträge für Lexeme, die auch in den Beispieltexten vorkommen. Eine größere lexikographische Unternehmung hätte den Rahmen dieser Arbeit gesprengt und wäre wenig sinnbringend gewesen, zumal GermaNet diese Lücke bereits schließt.

Zwei Einträge aus dieser semantischen Datenbank sehen dann z.B. wie folgt aus:

Netz {*Bahnlinie,Einkaufstasche,Fischnetz,Geflecht,Gewebe,*
Drahtgeflecht,Drahtnetz,Einkaufsnetz,Fangnetz,Fischernetz,
Fischnetz,Fischreuse,Flechtwerk,Gesamtheit,Gewebe,Haarnetz,
Maschenwerk,Netzwerk,Netzwerk,Reuse,System,Verknotung, Verschlingung,Wurfnetz,}

Ausgabe {*Auflage,Edition,Zuteilung,Unkosten,*}

Diese semantische Datenbank kann über die Menüleiste (Extras|SemDB laden) im Programm aufgerufen, editiert und erweitert werden. Dadurch ist zumindest manuell eine gewisse Flexibilität gewährleistet, wenn die eigentliche Dynamik im Sinne einer automatischen Pflege der Relationseinträge noch aussteht. Für eine solche dynamische Semantik wäre eine Gewichtung der einzelnen Relationen vonnöten. Barzilay und Elhadad versuchen genau diesen Effekt zu erzielen, indem sie die Wertung der bei der Topikkettenbildung beteiligten Relationen über Erfahrungswerte festlegen[19].

[16] nicht zu verwechseln mit der gleichnamigen Datenbank aus dem Verbmobil-Projekt
[17] (Mül97,)
[18] http://wortschatz.uni-leipzig.de/index.html, (26.05.2003)
[19] So erhalten z.B. Reiterationen und Synonyme einen Wert von 10, Antonyme 7 und Hyperonyme 4 (IM99, S. 114)

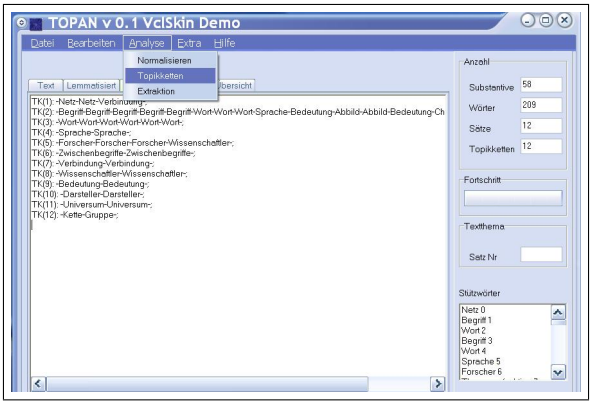


Abbildung 6.7: Analysevorgang starten

Denkbar wäre auch, diese Gewichtung der Einträge über die Häufigkeit ihres Auftretens in zu untersuchenden Texten zu steuern. Damit würde sich die semantische Datenbank den nutzerspezifischen Fachtexten von selbst anpassen.

6.3 Topikketten Analyse

In der Topikketten-Analyse besteht der eigentliche Kern des Programms. Die Voraussetzung für den Aufbau von isotopen Ketten lieferte die semantische Datenbank. Über die Menüleiste Analyse|Topikketten wird der Analysevorgang vom Nutzer gestartet[20].

Dieser Analysevorgang besteht aus folgenden Einzelschritten:

- Nachdem für alle Substantive des Zieltextes Einträge aus der semantischen Datenbank aufgesucht und in einen temporären Array geschrieben wurden,
- vergleicht TOPAN die einzelnen Einträge untereinander und
- sucht Reiterationen der Substantive im Text
- Eine Kette wird initialisiert, sobald ein Substantiv mehr als einmal im Text vorkommt.
- Bildet dieses Substantiv eine lexikalisch-semantische Relation zu einem anderen Substantiv (Vergleich der SemDB-Einträge), so wird dieses Substantiv der Kette hinzugefügt.
- Besteht eine Beziehung zwischen zwei Substantiven, die über keine Reiteration verfügen, wird ebenfalls eine neue Kette instanziiert.

[20] Siehe Abbildung 6.7, S. 62

Dabei kann es durchaus vorkommen, dass ein Substantiv in zwei unterschiedlichen Ketten auftritt: Wenn es zu mehren anderen Lexemen in Relation steht, wird es in eine weitere Kette eingefügt. Dieser Vorgang deutet die Ambiguität von Sememen an. Das so entstehende Problem einer Referenzidentität, auf das in Kapitel 4.1 hingewiesen wurde, taucht an dieser Stelle wieder auf. Ohne ein Hintergrundwissen ist diese Ambiguität vom Programm nicht zu lösen. Dieses Phänomen entspricht in etwa der mitunter sehr freien Assoziationstätigkeit beim Rezipieren von Texten, wenn gewisse Lexeme vom Rezipienten zusammengezogen werden, die vor dem Kontext des Textes nicht zusammengehören. Erscheinen in einem Text die Lexeme *Tasse* und *Fliegen* so könnte ein Gedankengang sicher Erinnerungen an den Themenbereich *fliegender Untertassen* hervorrufen, auch wenn der Text über die neueste Ausstattung in Flugzeugen berichtet und die eigentliche Topikkette für das Lexem *Tasse* in diesem speziellen Fall mit *Geschirr* in Verbindung gebracht werden müsste.

Inwieweit sich diese Ansammlung von disambiguierten Topikketten als störend bei die Ermittlung des Textthemas erweist, müßte eingehender erforscht werden. Zumindest hält sich die Beeinträchtigung schon allein deshalb in Grenzen, weil solche "ungewollten" Themen im Text seltener vorkommen als die vom Emittenten beabsichtigten. Daher werden die wesentlichen Topikketten i. A. länger als die "assoziierten" sein. Da die Themasatz-Extraktion auf der Grundlage der längsten Topikketten geschieht, kann eine Störung vermieden werden.

In der vorliegenden Arbeit ist dieser Programmabschnitt auf die selbstdefinierte semantische Matrix (SemDB) zugeschnitten. Eine künftige Weiterentwicklung müßte allerdings dafür Sorge tragen, dass ein solch umfassender Thesaurus wie GermaNet eingebunden werden kann[21], damit die angestrebte Domänunabhängigkeit garantieren wird.

Eine Weiterentwicklung von TOPAN müßte dem Nutzer zudem erlauben, den Schwellenwert von Übereinstimmungen in den Relationsbeziehungen zwischen zwei Lexemen – den eigenen Erfordernissen entsprechend – einzustellen. Damit könnte bei der Anwendung des Programmes aktiv auf die Qualität der zu bildenden Topikketten Einfluß genommen werden. Insbesondere bei der erwähnten semantischen Datenbank mit Klassenhierarchie wäre eine solche Feinabstimmung wünschenswert.

6.4 Satzextraktion

Die Satzextraktion schließt sich unmittelbar der Topikkettenbildung an (Analyse|Satzextraktion). In der vorliegenden Form ermittelt TOPAN den Themasatz bzw.

[21] Eine entsprechende C-Programmbibliothek (libwng.a) sieht der Lieferumfang von GermaNet vor: http://www.sfs.nphil.uni-tuebingen.de/lsd/, (26.05.2003)

Abbildung 6.8: Markierte Textthemasätze

die Themasätze über die Schnittmenge der Haupttopikketten. Nachdem das Programm die längsten Topikketten herausgefiltert hat, werden die einzelnen Texteme darauf überprüft, ob sie diese Topikketten enthalten. Von den Sätzen, bei denen dies zutrifft, ist nach dem *Agricolaschen Satz* davon auszugehen, dass sie wesentlich zur Bedeutung und Thematik des jeweiligen Textes beitragen und deshalb als Kernsätze des Textes anzusehen sind. Diese Ergebnissätze werden anschließend im Textfeld rot markiert und in einer Editbox numerisch ausgegeben[22].

Dieser Extraktionsalgorithmus schöpft jedoch keinewegs alle Möglichkeiten des theoretischen Modells aus. Die bestehende Programmversion unterstützt derzeit (noch) keine Regulierung der Kompression. Über eine Liste z.B. sollte der Nutzer aber die Wahl haben anzugeben, wieviele Sätze seine Zusammenfassung beinhalten muss. Erreicht werden könnte dies dadurch, dass der Text nach Angabe eines solchen Schwellenwertes die einzelnen Texteme solange auf Haupttopikketten überprüft, bis der gewünschte Kompressionsgrad erreicht ist.

Das Ermitteln von Kerninhalten eines Textes bleibt bei weitem nicht auf die Bestimmung der Schnittmengen von Topikketten und Sätzen beschränkt. Die von Buchbinder und Rozanov aufgestellten und in 4.2 behandelten Koeffizienten liefern einen mathematischen (wenn auch nicht absoluten[23])Bewertungsmaßstab für die semantische Abhängigkeit und Verflechtung der einzelnen Texteme.

Insbesondere der Koeffizient der semantischen Belastung und der lokale Koeffizient, deren Verlauf in Abbildung an Werten aus dem Beispieltext[24] dargestellt[25] ist, machen den Anteil der einzelnen Sätze am thematischen Schwerpunkt des Text augenscheinlich.

[22] Siehe Abbildung 6.8 S. 64

[23] Da sich die Koeeffizienten auf die aufgestellten Topikketten beziehen, deren Anordnung Schwankungen unterworfen ist, liefern sie nur Vergleichswerte.

[24] Siehe S. 79

[25] Siehe S. 84

Daraus ergibt sich zwingend die Notwendigkeit, die Berechnung des lokalen Koeffizienten in eine Folgeversion von TOPAN einzubauen. Dies war dem Verfassen aus Zeitgründen bei der aktuellen Referenimplementation nicht mehr möglich. Insbesondere die Anschaulichkeit eines thematischen Verlaufs im Text würde jedoch bei Auswahlprozessen den Nutzer eines Programms bei der automatisierten Textzusammenfassung hilfreich unterstützen.

7 Evaluation

7.1 Testreihe

In seinem Standardwerk zu "automatic summarization" weist MANI darauf hin[1], dass es bislang keine allgemein anerkannten Standards für die Evaluation von Programmen gibt, die Texte automatisch zusammenfassen.

So ist es schwierig im einzelnen abzuschätzen, wie eine "gute" Zusammenfassung letztlich auszusehen hat, da selbst professionelle Übersetzer und Journalisten ein und denselben Text unterschiedlich zusammenfassen. Hier könnte eine empirische Studie helfen: Einer größeren Anzahl von Testpersonen werden Texte vorgelegt, zu denen diese entweder Kurzfassungen schreiben sollen, oder aus denen sie wesentliche Passagen auswählen sollten. Auf diese Art und Weise könnte man eine generell akzeptabel Zusammenfassung (künstlich) kreieren, in deren Nähe sich auch das maschinell erstellte Kondensat befinden müßte, um einem gewissen "objektiven" Anspruch gerecht zu werden[2].

Ein weiteres Problem bildet der Kompressionsgrad: TOPAN wurde bisher nur für Texte mit einer Länge von ungefähr $200-300$ Wörtern getestet. Es hat sich jedoch gezeigt, dass die Qualität verschiedener Textanalyseverfahren bei längeren bzw. kürzeren Texten teils zu erheblichen Mängeln führen kann. Längere Texte könnten aber durchaus zerlegt werden und einzelne Passagen wiederum wie Teiltexte behandelt werden.

TOPAN setzt zudem (im gegenwärtigen Stadium) eine gewisse Datenbankpflege voraus. Die semantische Datenbank, wenn sie nicht von Drittanbietern (wie GermaNet)

[1] (IM99, S. 221)
[2] (IM99, S. 222)

bezogen wird, muß regelmäßig erweitert werden. In Einzelfällen müßten auch die automatisch generierten Topikketten manuell "nachjustiert" werden oder Schwellenwerte neu vergeben werden.

Bei den eigentlichen Tests von TOPAN an ausgewählten Beispieltexten[3] bildete das Programm – erwartungsgemäß – nicht dieselben Topikketten wie der Verfasser. Dies liegt ganz einfach im Aufbau der SemDB begründet: Die vielfachen Relationen zwischen Lexemeinträgen wachsen selbst bei einem kleinen Thesaurus ins Unüberschaubare. Umso interessanter war die Auswertung der elektronisch generierten Topikketten. In Abbildung 6.6 und 6.7 ist dass Ergebnis der Topikketten-Analyse und Satzextraktion am Beispieltext[4] zu sehen. Das Ergebnis ist ermutigend: Während die manuelle Analyse die Texteme T_4, T_9 und T_12 favorisierte, wählt TOPAN die Texteme T_4, T_9 und T_11 aus. Bei einem kleinen Text wie dem behandelten, gehören diese maschinell gewählten Sätze damit zu den wesentlichen Sätzen. Der Unterschied erklärt sich dadurch, dass TOPAN den Textaktanten *Sprache* zur Topikkette *Wort* hinzurechnet (ein legitimer Schritt) und dadurch vor einem leicht veränderten Verteilungsverhältnis steht. Dieses Ergebnis und der Rahmen in dem sich die Abweichungen bewegen, ist vergleichbar mit den unterschiedlichen Ergebnissen zu denen zwei Rezipienten an ein und demselben Text kommen würden. Dennoch sind die Unterschiede nicht willkürlich: TOPAN "assoziiert" seiner semantischen Datenbank entsprechend und folgt damit einem vorgegebenen Relationsschemata zwischen zwei Lexemen, sei es auf Basis von lexikalischen Beziehungen wie im vorliegenden Fall oder auf (minimal)distinktiven Semklassifikationen[5].

Ein erstaunliches Ergebnis bei der Anwendung von TOPAN, dass gleichzeitig die Stärke des Isotopiemodells verdeutlicht, soll an dieser Stelle nicht verschwiegen werden: So wurde in einem der Testläufe ein Text herangezogen, der zwar thematisch ("Sprache") in den Rahmen der in der SemDB hauptsächlich aufgenommenen Einträge fiel, für den der Verfasser allerdings mit Absicht keine eigenen Einträge in die Datenbank aufnahm. Dabei spielte die Fragestellung eine Rolle, inwieweit TOPAN Texte "zusammenfasst" bzw. zumindest analysiert, für die er nicht oder kaum auf die semantische Datenbank zurückgreifen kann. Das Ergebnis dieses Testlaufes

[3] Siehe Anhang C, S.87
[4] Siehe S. 79
[5] Vgl. das Besprochene in Kapitel 6.2, S. 58 und (Küh79, S. 81)

ist in Abbildung B.3, S. 86 nachgewiesen. Erstaunlicherweise hielt das Programm zumindest folgenden Satz für wesentlich:[6]

> "Das Ergebnis der spanischen Forscher: Eine Verteilung der Worthäufigkeiten, die dem Zipfschen Gesetz folgt, ergibt sich nur für Sprachen, die dem Prinzip der geringsten Anstrengung genügen, die also möglichst wenig Anstrengung von Redner und Zuhörer verlangen."[7]

Die Markierung der Schlüsselsätze funktionierte bei Testläufen auf einigen Rechnern nicht oder nur unzureichend. Aus diesem Grund wurde der GUI ein Editfeld hinzugefügt, dass zusätzlich darüber informiert, welche Sätze vom Textanalyse-Algorithmus ausgewählt wurden.

7.2 Ungelöste Probleme

Im Folgenden möchte ich rückblickend noch einmal die wesentlichsten Probleme und Schwierigkeiten zusammenfassen sowie notwendige Erweiterungen der Software andeuten:

- Ein eingehender Vergleich zwischen den verschiedenen Organisationsformen für semantische Datenbanken – insbesondere in Hinblick auf die Erfordernisse bei der Topikkettenbildung – muß erarbeitet werden.
- Die von Buchbinder und Rozanov aufgestellten Koeffizienten (K_s und K_l) zur Textisotopie müssen – ähnlich der Extraktion über Schnittmengenbildung – in ihrer Relevanz für die Textzusammenfassung näher untersucht werden.
- Der regressive Koeffizient muß vor dem Hintergrund des "sentence smoothing"[8] einer Evaluierung unterzogen werden. Mit ihm ließe sich vielleicht die Qualität der Satzextraktion bei der Textthemagewinnung (bei Anwendungen auf der Grundlage von Topikketten) deutlich steigern.
- Ein weitere Aufgabe besteht darin, die auf der Grundlage von Topikketten abstrahierte Kernbegrifflichkeit des Textes für automatische Textgenerierung (*abstracts*) heranzuziehen

[6] wobei die Stärke der Topikketten hier aufgrund mangelnder Einträge in der SemDB vor allem durch Reiterationen der im Text vorhanden Substantive bestimmt wurde und sich damit den durchaus passablen Ergebnissen anfänglicher automatisierter Textkondensationsprogramme anschließt.

[7] Siehe Anhang C, Textkorpus/Unit11

[8] Vgl. Kapitel 4.2, S.60

- Es wäre außerdem wünschenswert den Ansatz der Topikketten auf seine Relevanz beim Zusammenfassen unterschiedlicher Textsorten auszutesten. Eine Vermutung wäre, dass der Algorithmus insbesondere bei Zeitungsmeldungen und Artikeln aber auch noch Fachtexten greifen wird, während er bei Dialogen oder lyrischen Texten, die stark auf außersprachliche Realitäten verweisen, kaum brauchbare Ergebnisse liefern wird.

Abschließend läßt sich festhalten, dass eine Textthema-Ermittlung auf der Grundlage von Topikketten derart gute Ergebnisse liefert, dass ihr Einbinden in bestehende automatische Textkondensationsverfahren sowie ihre eigenständige Weiterentwicklung als durchaus lohnenswert erscheinen.

Vor diesem Hintergrund müßte das Isotopiemodell weiter ausgelotet werden. Es wäre denkbar, dass eine Textanalyse mittels Topikketten nicht allein auf das automatische Textzusammenfassen beschränkt bleiben wird, sondern auch in anderen Bereichen der Computerlinguistik erfolgversprechend eingesetzt werden könnte.

A Quellcode (Auszug)

```
//////////////////////////////////////////////////////////////
////Funktion: Substantivextraktion aus der Lemmatadatei
//////////////////////////////////////////////////////////////

void __fastcall TTopanMain::Normalisieren1Click(TObject *Sender)
{
int zeilen_anzahl = 0;
int i, temp, k = 0;
int a = 1;
int leerzeichen = 0;
int StartPos, ToEnd, FoundAt, FoundAtAlt;
bool nochwort;
TopanMain->satznr = 1;
TopanMain->wortnr = 1;
TopanMain->substnr = 0;

zeilen_anzahl = RichEdit2->Lines->Count; //read anzahl der zeilen

  for(i = 1; i <= zeilen_anzahl; i++){

    if(RichEdit2->Lines->Strings[i].AnsiPos(". SZE")){
      TopanMain->satzanfang[TopanMain->satznr] =
          RichEdit2->Lines->Strings[i+2];
```

```
    TopanMain->satznr++;
    a = 1;
  }
  if(RichEdit2->Lines->Strings[i].AnsiPos("@")){

    if(RichEdit2->Lines->Strings[i+2].AnsiPos("SUB")){
    //Wenn in der zweiten Zeile ein SUB
    //nach einem @ vorkommt, dann ist das Wort ein Subst.
    //deshalb wird das Wort aufgenommen:
    TopanMain->subst[TopanMain->substnr] =
        RichEdit2->Lines->Strings[i+1];
    //und zwar einmal in der Form, wie es im Text auftritt
    leerzeichen = RichEdit2->Lines->Strings[i+2].AnsiPos("␣");
    //einmal in Lemmataform
    //dazu wird position des ersten Leerzeichens bestimmt
    //um dann das Lemmata zu extrahieren
    TopanMain->satzwort[TopanMain->satznr][a] =
        RichEdit2->Lines->Strings[i+2].SubString(1, (leerzeichen-1));
    a++;
    ////////////////////////
    if(!RichEdit2->Lines->Strings[i+2].AnsiPos("?")){
    TopanMain->lemma[TopanMain->substnr] =
        RichEdit2->Lines->Strings[i+2].SubString(1, (leerzeichen-1));
    }
    else{TopanMain->lemma[TopanMain->substnr] =
        RichEdit2->Lines->Strings[i+1];}
    TopanMain->substnr++;
    //subst Zaehler Inkrement
    }
    if(! ((RichEdit2->Lines->Strings[i+2].AnsiPos("SZE")
    || (RichEdit2->Lines->Strings[i+2].AnsiPos("SZT"))
    || (RichEdit2->Lines->Strings[i+2].AnsiPos("SZK"))))){
    //Anzahl der Wortformen im Text zaehlen
    //ohne Satzzeichen
    TopanMain->wortnr++;
    }
  }
```

```
  }
  //Edit1->Text = TopanMain->satzwort[4][1];
  //Testausgabe
  SatzAnzahl->Text = IntToStr(TopanMain->satznr);

  SubstAnzahl->Text = IntToStr(TopanMain->substnr);
  //Ausgabe der Anzahl an Substantiven
  WrtAnzahl->Text = IntToStr(TopanMain->wortnr);
  //Ausgabe der Anzahl an Wortformen
  temp = (TopanMain->substnr - 1);
  //Hilfsvariable fuer Einfaerben
  Registerkarte->ActivePageIndex = 0;
  //Registerkarte Textfeld aktivieren


////////////////////////////////////////////////////
///Topikketten Analyse
////////////////////////////////////////////////////

void __fastcall TTopanMain::Topikketten1Click(TObject *Sender)
{

try{

int j, k = 1;
int n = 1;
int p, q, r, l, m, t, v, x = 0;
int rekurrenz = 0;
int treffer = 0;
int schwellenwert = 0;

bool nochglieder;
bool vorhanden;
bool schongefunden;

AnsiString kette;
AnsiString SemDB[200][30];
```

```
String S;

//////////////////////////////////////////
//Fortschrittsanzeige                   //
//////////////////////////////////////////

try{
TKProgressBar->Min = 0;
TKProgressBar->Max = (TopanMain->substnr - 1);
}
catch (...){
 Application->MessageBoxA(
 "Bitte zuerst Lemmata normalisieren", "Fehler",
 MB_ICONHAND | MB_OK);
}
//////////////////////////////////////////
//Öffne semantische Datenbank           //
//////////////////////////////////////////

//int fHandle = 0;
try{
     //iFileHandle = FileOpen("SemDB.txt", fmOpenRead);
     RichEdit4->Lines->LoadFromFile("c:/SemDB.txt");

}
catch (...){
 Application->MessageBoxA(
 "SemDB konnte nicht geoeffnet werden", "Fehler",
 MB_ICONHAND | MB_OK);
}

//////////////////////////////////////////
//Lese SemDB ein                        //
//////////////////////////////////////////
TStringList* lines = new TStringList;
lines->LoadFromFile("c:/SemDB.txt");
```

```
for(r = 0; r < TopanMain->substnr; r++){
    ListBox1->Items->Add(TopanMain->lemma[r] + "␣" + r);
    l = 0;
    for(; l < lines->Count; l++){

       S = lines->Strings[l];
       //lese zeile in temp variable ein
       int pos = S.Pos("{");
       String Eintrag = S.SubString(1, (pos-2));
       S.Delete(1, pos);
       //lese SemDB Eintrag

        if(AnsiStrComp(Eintrag.c_str(),
                TopanMain->lemma[r].c_str()) == 0){
        //
                //jetzt beginnt die eigentliche arbeit,
                //das auslesen der synrelationen
        //
                for(x = 0; x <30; x++){
        pos = S.Pos(",");

         if(pos!=0){
         SemDB[r][x] = S.SubString(1, pos-1);
         S.Delete(1, pos);
         }
         }
      l = (lines->Count-1);
     }
    }
}

//////////////////////////////////////
//Bilde Topikketten                 //
//////////////////////////////////////

for(j = 0; j < TopanMain->substnr; j++){
   for(k = (j+1); k < TopanMain->substnr; k++){
```

```
            if((TopanMain->lemma[j] != "") &&
            (TopanMain->lemma[k] != "")){

              ////////////////////////////////Rekurrenzvergleich

              if( ((TopanMain->lemma[j]) == (TopanMain->lemma[k]))
              && (rekurrenz <1 )){
              treffer++;
                  if(treffer>=schwellenwert){
                                //
                  //Treffer!Das Wort kommt (1.mal) nochmal im Text vor
                  m++;
                                //
                  //neue kette!
                  TopanMain->topik[m][n] = TopanMain->lemma[j];
                  TopanMain->topik[m][++n] = TopanMain->lemma[k];
                  rekurrenz = 1;
                  TopanMain->tklang[m] = n;
                  //trat mehr als einmal auf, kette bereits init.
                  TopanMain->lemma[k] = "";
                                //
                  //Eintrag loeschen - verhindert Mehrfachnennung
                  }
              }
              if (((TopanMain->lemma[j]) == (TopanMain->lemma[k]))
              && (rekurrenz >=1)){
              treffer++;
                if(treffer>=schwellenwert){
                //Treffer! Wort kommt mehrmals im Text vor
                TopanMain->topik[m][++n] = TopanMain->lemma[k];
                TopanMain->lemma[k] = "";
                }
              }
            }
}
n = 1;
```

```
    rekurrenz = 0;
    treffer = 0;

    TKProgressBar->StepBy(1);
    //Erhoehe die Fortschrittsanzeige um 1 (Satz)
}


////////////////////////////////////////////////////////
///Themasatz Extraktion
////////////////////////////////////////////////////////

void __fastcall TTopanMain::Extraktion1Click(TObject *Sender)
{
int i = 0;
int k, t, j = 0;
int ketten_anzahl = 0;
String S;
bool end = true;

ketten_anzahl = RichEdit3->Lines->Count;
//read anzahl der zeilen

////////////////////////////////////////////////////
///Aufnehmen der neuesten Topikketten            ///
////////////////////////////////////////////////////

for(i = 0; i < ketten_anzahl; i++){

    S = RichEdit3->Lines->Strings[i];
    if(S != ""){
            //lese kettenzeile in temp variable ein
            int pos = S.Pos("-");
            S.Delete(1, pos);
            //TK Nr loeschen
            while(end){
                if(S.Pos(";") == 1){
```

```
                end = false;
                }
                int pos = S.Pos("-");
                String glied = S.SubString(1, (pos-1));
                TopanMain->topik[i][++j] = glied;
                S.Delete(1, pos);
            }
            end = true;
            TopanMain->tklang[i] = (j-1);
            j = 0;
    }
}

//////////////////////////////////////////////////////
///Schnittmenge der HauptTopikketten                ///
//////////////////////////////////////////////////////

int x, y, z, d = 0;
int themasatz[10];

...

for(x = 0; x < TopanMain->satznr; x++){
 for(z=0; z<20;z++){
    for(y=0; y<40; y++){
     if(( TopanMain->satzwort[x][y] ==
          TopanMain->topik[haupttk][z])
     && (TopanMain->topik[drittlangste][z] != "")){
     themasatz[++d] = x;
     RichEdit3->Lines->Add(IntToStr(themasatz[d]));
     z= 19 ;
     y= 39 ;
     }
     }
 }
}
...
```

B Textbeispiel und Tabellen

Quelle: Bild der Wissenschaften[1]

(T1)Das Netz der Begriffe ist eng geknüpft

(T2)Im Englischen sind alle Wörter über zwei Begriffe miteinander "verwandt"

(T3)Wie eng die meisten Wörter einer Sprache miteinander verbunden sind, haben nun amerikanische Forscher gezeigt:

(T4)Über die Thesaurus-Funktion eines gewöhnlichen Textverarbeitungsprogramms lässt sich über nur zwei Zwischenbegriffe eine Verbindung zwischen beliebigen Wörtern herstellen, meldet das Fachmagazin "Nature" in seiner Online-Ausgabe.

(T5)Die Wissenschaftler um Adilson Motter von der Universität in Tempe (USA) hatten 30.000 englische Vokabeln aus einem Online-Thesaurus ausgesucht.

(T6)Solche Programme helfen Autoren beim Suchen nach Wörtern mit gleicher oder ähnlicher Bedeutung.

(T7)Aus diesen Begriffen wählten die Forscher nach dem Zufallsprinzip beliebige Paare aus und suchten nach der kürzesten Verbindung zwischen den beiden.

(T8)Von "Darsteller" zu "Universum" beispielsweise gelangten sie über die Kette: Darsteller – Charakter – Natur – Universum.

(T9)Meistens reichten zwei Zwischenbegriffe aus, um zwei beliebige Wörter miteinander zu verbinden, berichten die Wissenschaftler.

(T10)Allerdings hatten sie für die Untersuchung nur "Wurzelwörter" ausgesucht, für die im Thesaurus auch verwandte Begriffe angegeben sind.

[1] http://www.wissenschaft.de/sixcms/detail.php?id=149211

(T11)Die Ergebnisse seien auch auf andere Sprachen übertragbar, vermuten die Forscher.

(T12)Das eng geknüpfte Netz ähnlicher oder verwandter Wörter sei wohl ein Abbild der Strukturen in unserem Gehirn, das gleichfalls Begriffe mit ähnlicher Bedeutung in Gruppen zusammenfasst und speichert.

Textem [a]	Tk_1	Tk_2	Tk_3	Tk_4	Tk_5	Tk_6	Tk_7	Tk_8
T_1	Netz				Begriff			
T_2	‡	Wörter			Begriffe			
T_3	‡	Wörter	Sprache	Forscher	‡			
T_4	Verbindung	Wörter	‡	‡	Zwischenbegriffe	Thesaurus	Programm	
T_5	‡	Vokabel	‡	Wissenschaftler	‡	Thesaurus	‡	
T_6	‡	Wörter	‡	‡	‡	‡	Programme	Bedeutung
T_7	Verbindung	‡	‡	Forscher	Begriff	‡		‡
T_8	Kette	‡	‡	‡	‡	‡		‡
T_9	‡	Wörter	‡	Wissenschaftler	Zwischenbegriffe	‡		‡
T_{10}	‡	Wurzelwörter	‡		Begriffe	Thesaurus		‡
T_{11}	‡	‡	Sprache		‡			‡
T_{12}	Netz	Wörter			Begriffe			Bedeutung

Tabelle B.1: Übersicht der Topikketten

[a] Schwellenwert: Lexemanzahl ≥ 2

Textem	K_s	K_l	$HT_{k3} \cap HT_{k6}$	Abweichung vom mittleren K_l
T_1	0,25	0,176	–	+0,005
T_2	0,13	0,221	$\surd$	+0,050
T_3	0,214	0,163	–	$-0,008$
T_4	0,2	0,23	$\surd$	$\underline{+0,059}$
T_5	0,2	0,165	–	$-0,006$
T_6	0,2	0,125	–	$-0,046$
T_7	0,188	0,197	–	+0,026
T_8	0,2	0,083	–	$-0,088$
T_9	0,158	0,227	$\surd$	$\underline{+0,056}$
T_{10}	0,16	0,213	$\surd$	+0,042
T_{11}	0,5	0,023	–	$-0,148$
T_{12}	0,15	0,23	$\surd$	$\underline{+0,059}$

Tabelle B.2: Semantische Belastung und Mittelwert aus den lokalen Koeffizienten gegenübergestellt

Textem	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}	T_{11}	T_{12}
T_1	–	0,25	0	0,286	0	0	0,4	0,3	0,2	0,2	0	0,3
T_2	0,25	–	0,2	0,286	0,2	0,2	0,2	0	0,4	0,4	0	0,3
T_3	0	0,2	–	0,125	0,3	0,16	0,16	0	0,3	0,16	0,25	0,143
T_4	0,286	0,286	0,125	–	0,25	0,25	0,25	0,16	0,25	0,375	0	0,3
T_5	0	0,2	0,3	0,25	–	0,16	0,16	0	0,3	0,3	0	0,143
T_6	0	0,2	0,16	0,25	0,16	–	0	0	0,16	0,16	0	0,286
T_7	0,4	0,2	0,16	0,25	0,16	0	–	0,25	0,3	0,16	0	0,286
T_8	0,3	0	0	0,16	0	0	0,25	–	0	0	0	0,2
T_9	0,2	0,4	0,3	0,25	0,3	0,16	0,3	0	–	0,3	0	0,286
T_{10}	0,2	0,4	0,16	0,375	0,3	0,16	0,16	0	0,3	–	0	0,286
T_{11}	0	0	0,25	0	0	0	0	0	0	0	–	0
T_{12}	0,3	0,3	0,143	0,3	0,143	0,286	0,286	0,2	0,286	0,286	0	–

Tabelle B.3: Lokaler Koeffizient – Maß für die isotope Verflechtung zwischen Textemen

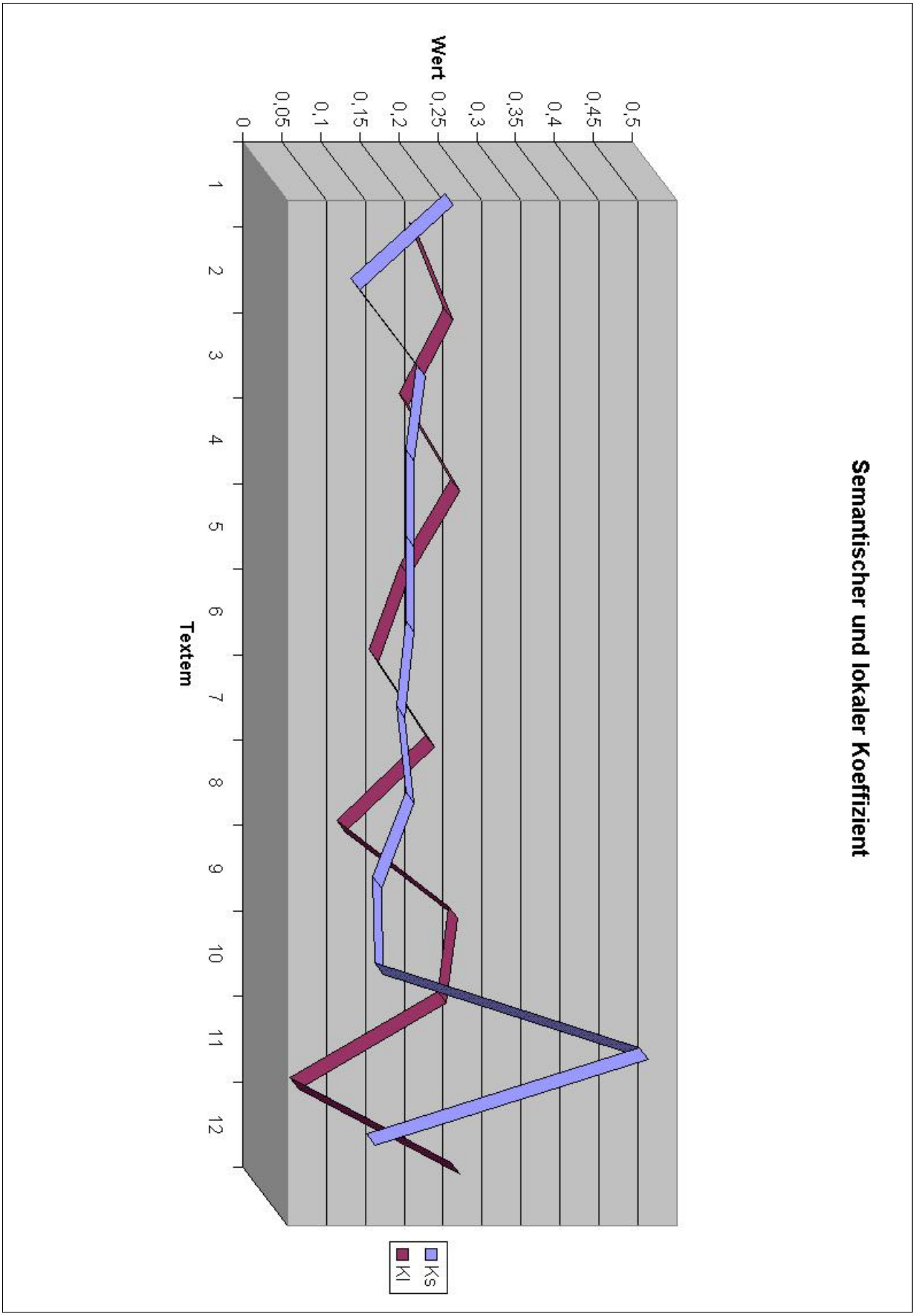


Abbildung B.1: Semantischer und lokaler Koeffizient im Vergleich – Diagramm

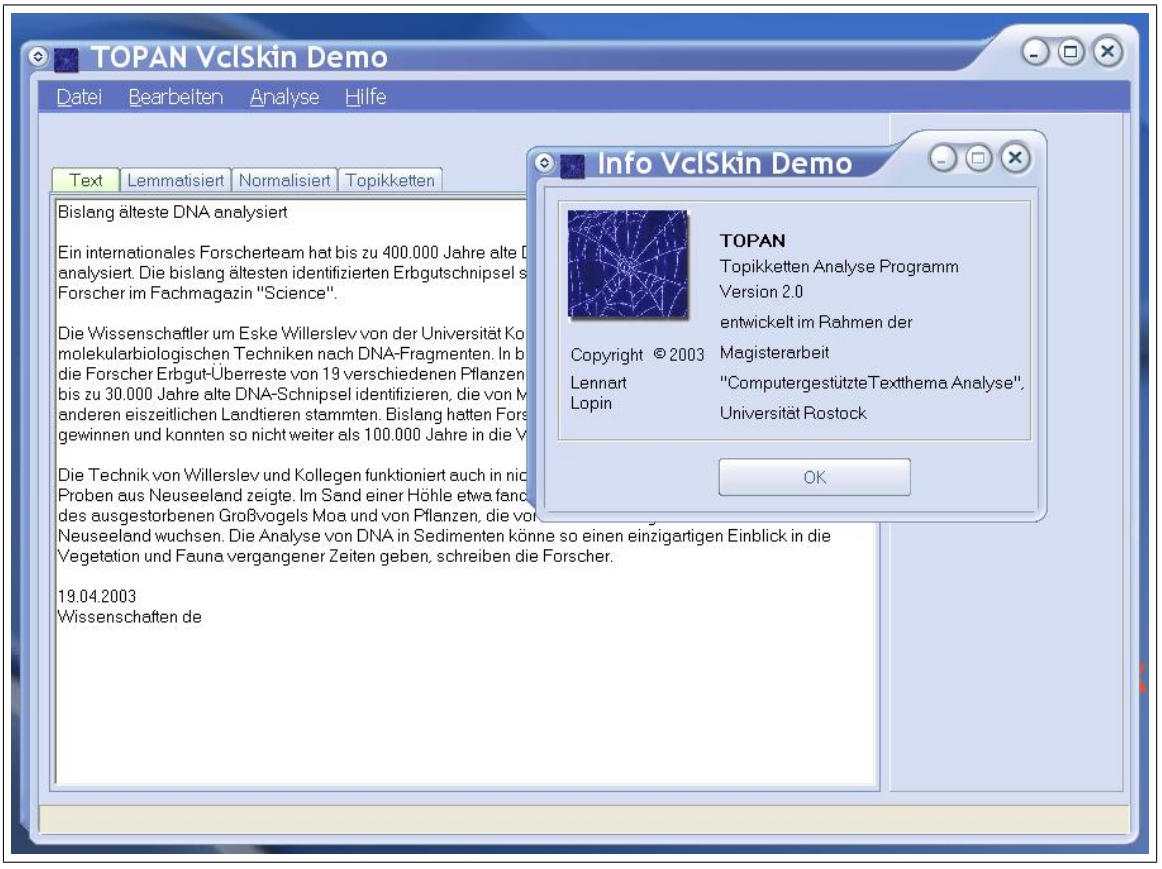


Abbildung B.2: Topan

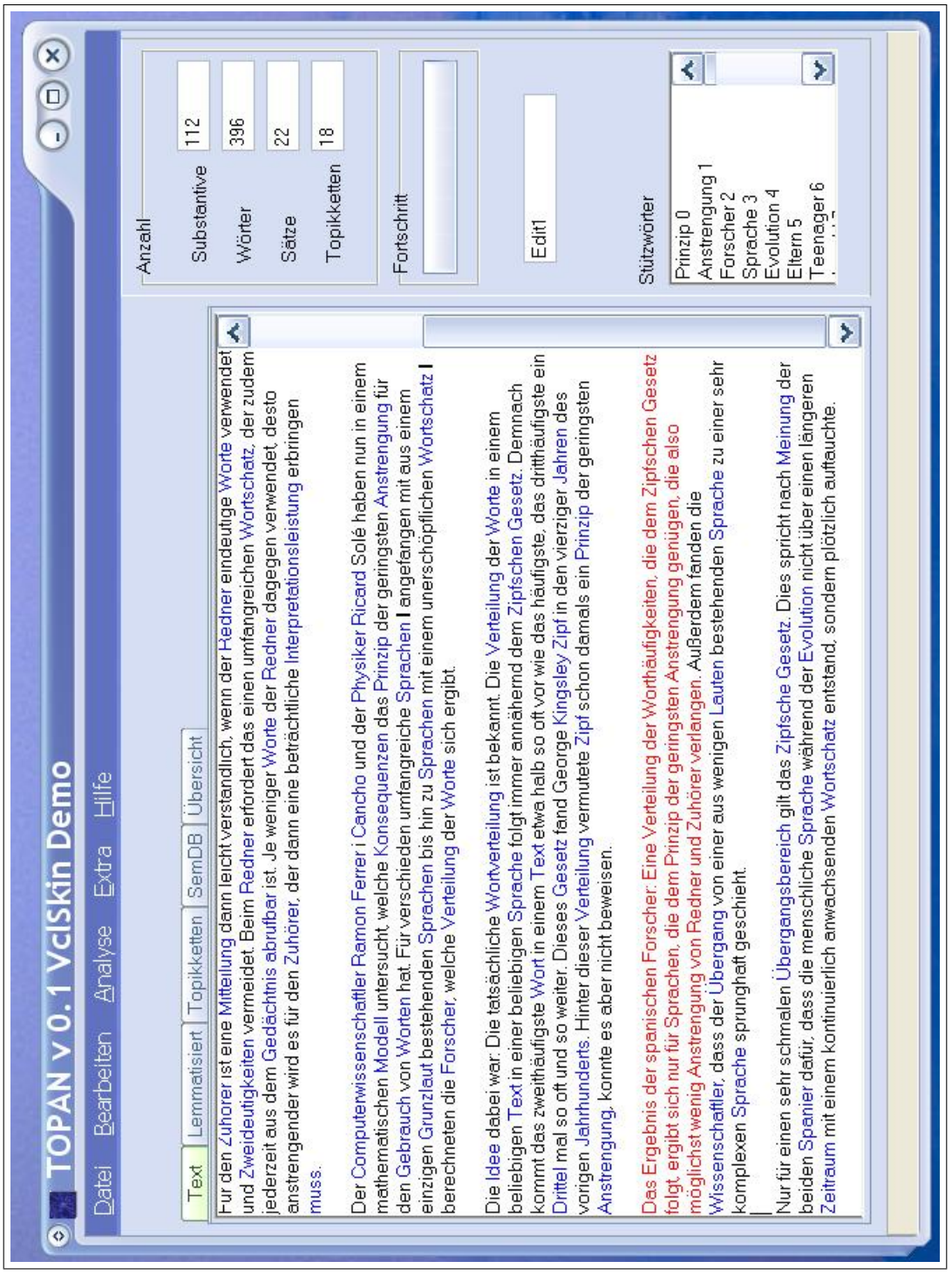


Abbildung B.3: Themasatz: Ermittelt an einem Beispieltext ohne spezifische Einträge in der SemDB

C Inhalt der CDROM

Auf der beiliegenden CDROM finden sich folgende ergänzende Materialen:

Dokumentation In diesem Ordner befindet sich eine elektronische Version der vorliegenden Magisterarbeit im pdf-Dateiformat.

Artikel In diesem Verzeichnis befindet sich eine Zusammenstellung von akademischen Online-Publikationen, auf die zum Teil im Rahmen der Magisterarbeit Bezug genommen wurde.

Software Dieser Ordner enthält die aktuellste Version von TOPAN samt Installationsanleitung. Der morphologische Tagger Morphy liegt ebenfalls bei.

Textkorpus Eine Reihe von Beispieltexten sind in diesem Ordner enthalten. Für einige dieser Texte bestehen bereits Einträge in der semantischen Datenbank (letztere ist im Software-Ordner zu finden). Der in der Arbeit behandelte Beispieltext (Anhang B) ist darin unter der Bezeichnung “Unit3” mitinbegriffen.

lepage

Literaturverzeichnis

[AB01] A. Budanitsky, G. Hirst: *The Verbmobil Semantic Database.* (12.05.2003), Juni 2001. <perso.wanadoo.fr/heinecke/litlist/doc/konvens.ps.gz>.

[AEM02] Adilson E. Motter, Alessandro P. S. de Moura, Ying-Cheng Lai Partha Dasgupta: *Topology of the conceptual network of language.* Physical Review E, 65(065102):1–4, April 2002.

[Agr79] Agricola, Erhard: *Textstruktur–Textanalyse–Informationskern.* VEB Verlag Enzyklopädie, Leipzig, 1979.

[AL01] Angelika Linke, Markus Nussbaumer, Paul R. Portmann: *Studienbuch Linguistik.* Nummer 121 in *Reihe Germanistische Linguistik.* Niemeyer, Tübingen, 2001.

[Ash94] Asher, R. E. (Herausgeber): *Encyclopedia of Language and Linguistics.* Pergamon Press, Oxford, 1994.

[BH01] Birgit Hamp, Helmut Feldweg: *GermaNet - a Lexical-Semantic Net for German.* www.ifi.unizh.ch/CL/hess/classes/seminare/semrep/docs/germanet.pdf, (12.05.2003), Juni 2001.

[BJ95] Byung-Jin, Choi: *Vererbungsbasierte semantische Repräsentation für maschinelle Wörterbücher.* Frankfurt am Main, 1995.

[Bri92] Brinker, Klaus: *Linguistische Textanalyse. Eine Einführung in Grundbegriffe und Methoden.* Berlin, 1992.

[FX01] Feiyu Xu, Melanie Siegel, Günter Neumann: *Customizing GermaNet for the Use in Deep Linguistic Processing.* http://www.dfki.de/, (12.05.2003), Juni 2001.

[Hei97] Heid, Ulrich: *Zur Strukturierung von einsprachigen und kontrastiven elektronischen Wörterbüchern.* Niemeyer, Tübingen, 1997.

[Hei00] Heinemann, Wolfgang: *Das Isotopiekonzept.* In: K. Brinker, G. Antos (u.a.) (Herausgeber): *Textlinguistik. Dialog*, Seiten 54–59. de Gruyter, Berlin, 2000.

[IM99] Inderjeet Mani, Mark T. Maybury (Herausgeber): *Advances in automatic text summarization.* MIT Press, Cambridge, Massachusetts, 1999.

[JAE30] J. A. Eberhard, J. G. E. Maasz, J. G. Gruber: *Versuch einer allgemeinen teutschen Synonymik in einem kritisch – philosophischen Wörterbuche der sinnverwandten Wörter der hochteutschen Mundart.* Halle, Dritte Auflage, 1826–1830.

[Kai00] Kaiser, Ulrich: *C/C++. Von den Grundlagen zur professionellen Programmierung.* Galileo Computing. Galileo, Bonn, 2000.

[Küh79] Kühn, Peter: *Der Grundwortschatz. Bestimmung und Systematisierung.* Nummer 17 in *Reihe Germanistische Linguistik.* Niemeyer, Tübingen, 1979.

[KW01] Karsten Worm, Johannes Heinecke: *The Verbmobil Semantic Database.* http://www.coli.uni-sb.de/ worm/work.html, Juni 2001.

[Löt78] Lötscher, Andreas: *Text und Thema.* Nummer 81 in *Reihe Germanistische Linguistik.* Niemeyer, Tübingen, 1978.

[Man01] Mani, Inderjeet: *Automatic summarization.* Nummer 3 in *Natural Language Processing.* John Benjamins, Philadelphia, 2001.

[Mar00] Marcu, Daniel: *The theory and practice of discourse parsing and summarization.* MIT, Cambridge, Massachusetts, 2000.

[MB01] Meru Brunn, Yllias Chali, Christopher J. Pinchak: *Text Summarization Using Lexical Chains.* In: *Document Understanding Conferences*, Gaithersburg, 2001. Department of Mathematics and Computer Science. University of Lethbridge., National Institute of Standards and Technology. Information Access Division.

[MH02] Margot Heinemann, Wolfgang Heinemann 2002: *Grundlagen der Textlinguistik. Interaktion – Text – Diskurs.* Nummer 230 in *Reihe Germanistische Linguistik.* Niemeyer, Tübingen, 2002.

[Mül97] Müller, Wolfgang (Herausgeber): *Sinn- und sachverwandte Wörter. Synonymwörterbuch der deutschen Sprache.* 8. Dudenverlag, Mannheim, Second Auflage, 1997.

[RB81] Robert Beaugrande, Wolfgang Dressler: *Einführung in die Textlinguistik.* Reihe Germanistische Linguistik. Niemeyer, Tübingen, 1981.

[Sch97] Schutte, Jürgen: *Einführung in die Literaturinterpretation.* Nummer 217 in *Sammlung Metzler.* J. B. Metzler, Stuttgart, Fourth Auflage, 1997.

[SIB89] S. István Bátori, Winfried Lenders, Wolfgang Puschke (Herausgeber): *Computerlinguistik. Ein internationales Handbuch zur computergestützten Sprachforschung und ihrer Anwendungen.* De Gruyter, Berlin, 1989.

[Som96] Sommerville, Ian: *Software Engineering.* International Computer Science Series. Addison-Wesley, New York, Fifth Auflage, 1996.

[SS88] Seppo Sippu, Eljas Soisalon-Soininen: *Parsing Theory. Languages and Parsing*, Band 1 der Reihe *EATCS.* Springer, Berlin, 1988.

[VAB75] V. A. Buchbinder, E. D. Rozanov: *O celostnosti i strukture teksta.* Voprosy Jazykoznanija, Moskva, 6:73–86, 1975.

[Web00] Weber, Nico: *Die Semantik von Bedeutungsexplikationen.* Nummer 3 in *Sprache, Sprechen und Computer.* Peter Lang, Frankfurt am Main, 2000.

[WH91] Wolfgang Heinemann, Dieter Viehweger: *Textlinguistik. Eine Einführung.* Nummer 115 in *Reihe Germanistische Linguistik.* Niemeyer, Tübingen, 1991.

[YAW96] Yorik A. Wilks, Brian M. Slator, Louise M. Guthrie: *Electric Words. Dictionaries, Computers, and Meanings.* MIT, Cambridge, Massachusetts, 1996.

Hiermit versichere ich, die Arbeit selbständig erstellt und keine anderen als die angegebenen Hilfsmittel benutzt zu haben.

Lennart Lopin