UNIVERSITÄT DORTMUND
FACHBEREICH INFORMATIK

LEHRSTUHL VIII
KÜNSTLICHE INTELLIGENZ

Automatische Kategorisierung von Volltexten unter Anwendung von NLP-Techniken

LS–8 Report 22

Sandra Schewe

Dortmund, 16.September 1997

Universität Dortmund
Fachbereich Informatik

University of Dortmund
Computer Science Department

Forschungsberichte des Lehrstuhls VIII (KI)
Fachbereich Informatik
der Universität Dortmund

ISSN 0943-4135

Anforderungen an:

Universität Dortmund
Fachbereich Informatik
Lehrstuhl VIII
D-44221 Dortmund

Research Reports of the unit no. VIII (AI)
Computer Science Department
of the University of Dortmund

ISSN 0943-4135

Requests to:

University of Dortmund
Fachbereich Informatik
Lehrstuhl VIII
D-44221 Dortmund

e-mail: reports@ls8.informatik.uni-dortmund.de
ftp: ftp-ai.informatik.uni-dortmund.de:pub/Reports
www: http://www-ai.informatik.uni-dortmund.de/ls8-reports.html

Automatische Kategorisierung von Volltexten unter Anwendung von NLP-Techniken

LS–8 Report 22

Sandra Schewe

Dortmund, 16.September 1997

Universität Dortmund
Fachbereich Informatik

Zusammenfassung

Die vorliegende Arbeit befaßt sich mit der Informationsgewinnung aus Daten, wie sie das World Wide Web zur Verfügung stellt. Dabei liegt der Schwerpunkt auf der Verarbeitung von Volltexten, denn ein großer Anteil der Daten ist im WWW in dieser Form verfügbar. Zur Unterstützung der Informationsgewinnung werden die Volltexte kategorisiert, so daß ein Benutzer entweder gezielt in einer Kategorie nach bestimmten Informationen suchen kann, oder so daß ihm nach Themen sortierte Texte vorgelegt werden können, aus denen er nach Interesse Themengebiete auswählen kann.

Zur Kategorisierung der Texte werden Techniken aus dem Bereich Natural Language Processing, kurz NLP-Techniken, herangezogen. Überlegungen zu den besonderen Eigenschaften der deutschen Sprache führen zu der hier vorgestellten Verfahrensweise. Experimente werden zeigen, in wie weit der Einsatz von NLP-Techniken und damit die Berücksichtigung von Sprache von Nutzen ist.

Inhaltsverzeichnis

1 Einleitung

> There may be more text data in electronic form than ever before, but much of it is ignored. No human can read, understand, and synthesize megabytes of text on an everyday basis. Missed information – and lost opportunities – has spurred researchers to explore various information management strategies to establish order in the text wilderness ([Cowie und Lehnert, 1996], S.80).

Weltweite Kommunikation und Informationsverbreitung und die technische Entwicklung dazu einsetzbarer Medien wie das Internet bieten einer täglich wachsenden Zahl Menschen die Möglichkeit, Informationen anzubieten und sich selbst zu informieren. Nicht nur Universitäten und Firmen, sondern auch private Haushalte nutzen dieses Angebot. Besonders das World Wide Web (WWW) wird als Plattform zur Informationverbreitung und als Informationsquelle immer populärer. Die Teilnahme so vieler verschiedener Institutionen und Privathaushalte zieht eine sehr große Bandbreite zum Angebot stehender Themen nach sich. Der Kreativität der Anbieter sind bei der Gestaltung ebenfalls keine Grenzen gesetzt. So präsentiert sich das WWW einem browsenden Benutzer in einer großen Gestaltungs- und Themenvielfalt.

Obwohl diese Größe und Vielfalt zu den Vorzügen des WWW zählt, bringt sie auch einen Nachteil mit sich: Die riesige Menge an zur Verfügung stehenden Daten ist unüberschaubar. Benutzer, die nach einer bestimmten Information suchen, sehen sich einer Datenflut gegenüber, die das Auffinden der gesuchten Information erschwert oder unmöglich macht.

Denn eine große Menge an Daten bedeutet nicht gleichzeitig auch eine große Menge an Information. Daten werden erst dann zu Information, wenn der Adressat sie auch nutzen kann, d.h. wenn er sie verwenden kann, weil er sie benötigt oder weil sie ihn interessieren. Nicht verwendbare Daten stellen somit auf der Suche nach Information eher ein Hindernis dar. Natürlich soll dieser Nachteil keine Einschränkung der a priori vorliegenden Datenmenge zur Folge haben, denn es liegt im subjektiven Ermessen jedes Benutzers, welche Daten für ihn zu Information werden.

Wie aus dem einleitenden Zitat deutlich wird, geht es also darum, den Benutzer bei der Gewinnung von Information aus den Datenmengen zu unterstützen. Dazu sind verschiedene Strategien denkbar. Eine Möglichkeit ist es, aus der gesamten Datenmenge eine Vorauswahl zu treffen und dem Benutzer gemäß seines Interesses diese kleineren Teilmengen vorzustellen. Andere Ansätze gehen noch weiter mit dem Versuch, Informationsbedürfnisse und Fragen direkt aufgrund der gegebenen Daten zu beantworten. Eine dritte mögliche Herangehensweise ist es, die Daten so zu strukturieren, daß die Suche nach Information erleichtert wird. Diese dritte Strategie bildet den Schwerpunkt dieser Arbeit.

Neben verschiedenen Strategien sind auch verschiedene Techniken zur Analyse der Daten anwendbar. Eine Gruppe von Techniken, die zur Verarbeitung von Textdaten eingesetzt wird, gehört in das Gebiet des *Natural Language Processing*. Hier wird die Frage untersucht, ob Techniken aus diesem Gebiet bei der Informationsgewinnung von Nutzen sein können.

Die folgenden beiden Abschnitte konkretisieren die Aufgabenstellung und geben eine Übersicht über den Aufbau dieser Arbeit.

1.1 Aufgabenstellung

Diese Arbeit befaßt sich mit einem Ausschnitt des World Wide Web: Bei den Daten, die verarbeitet werden sollen, handelt es sich um Zeitungsartikel aus on-line angebotenen Tageszeitungen. Auch wenn damit nur ein Ausschnitt des WWW gewählt wurde, sind diese zwei Bereiche in gewisser Weise vergleichbar. Beide bieten eine Vielfalt an Themen an, was nach sich zieht, daß nicht alle Themen für alle Anwender interessant sind. Um das Themenangebot der Zeitungsartikel übersichtlicher zu gestalten, sollen die als Volltexte vorliegenden Artikel kategorisiert werden, so daß dem Benutzer sofort alle Artikel zu einem Thema vorgestellt werden können.

Um das Thema eines Textes zu bestimmen, muß ein Mensch diesen Text verstehen. Sein Werkzeug dazu ist die Sprache. Es liegt also nahe, zur automatischen Bestimmung des Themas eines Textes die Sprache zu berücksichtigen. Dazu bieten sich die Techniken aus dem Gebiet Natural Language Processing (NLP) an. Diese unterscheiden sich darin, auf welcher Ebene der Sprache sie ansetzen, und auf welchem Teilgebiet der Linguistik sie basieren. Je nach Ebene und Teilgebiet erhöht sich der Aufwand, den die Verfahren betreiben müssen, um erfolgreich zu analysieren. Hinzu kommt, daß es sich bei der zu analysierenden Sprache um Deutsch handelt. Während bei flexionsarmen Sprachen wie Englisch weniger aufwendige Verfahren zum Einsatz kommen, muß im Deutschen ein größeres System an Regeln für den automatisierten Einsatz umgesetzt werden. Hier geht es also auch darum, bei der Wahl der Techniken aus dem Bereich NLP ein günstiges Verhältnis zwischen Aufwand und Nutzen zu finden. Experimente sollen zeigen, ob die Berücksichtigung von Sprache durch Einsatz von NLP-Techniken sinnvoll ist.

1.2 Übersicht

Das dieser Übersicht folgende Kapitel 2 befaßt sich mit dem Stand der Forschung. Dabei werden die Strategien zur Informationsgewinnung unter dem Aspekt der Anwendung von NLP-Techniken beleuchtet. Aus den Möglichkeiten und Grenzen auf dem Gebiet Natural Language Processing und aus den Besonderheiten der Domäne *Deutsche Tageszeitungen*, die Kapitel 3 schildert, leitet sich das Vorgehen im Rahmen dieser Arbeit ab. Das sich anschließende Kapitel 4 stellt die Bestandteile des Systems AKAT (Automatische KAtegorisierung von Texten) im Detail vor. Kapitel 5 erläutert schließlich die durchgeführten Experimente und beschreibt die Ergebnisse. Eine abschließende Zusammenfassung beendet diese Arbeit und wirft einen kurzen Blick auf mögliche Anwendungen.

2 Forschungsstand

Zwei Bereiche der Informationstechnik sind von der in der Einleitung angesprochenen Datenflut besonders betroffen: In dem Bemühen, möglichst viele Daten zu speichern, um aus ihnen weitere Informationen zu gewinnen, vervielfachten sich in den letzten Jahren die in Datenbanken gespeicherten Datenmengen (vgl. [Franzel, 1996]). Die zweite, sich täglich vergrößernde Informationsquelle ist das World Wide Web, das Daten zu vielen verschiedenen Themenbereichen zur Verfügung stellt. Die Aufgabe, Informationen aus den gegebenen Daten zu gewinnen, haben beide Forschungsgebiete gemeinsam. Den Unterschied bilden die Daten: Datenbanken sind streng strukturiert. Dem gegenüber steht die vernetzte Struktur des WWW, die keinen Regeln unterliegt. Hinzu kommt, daß die Einträge einer Datenbank eine definierte Bedeutung haben, während die Informationsanbieter des WWW keinen Beschränkungen bezüglich der Gestaltung ausgesetzt sind. Zur Problematik der Datenflut in Datenbanken sei auf das Gebiet der Wissensentdeckung in Datenbanken verwiesen. Hier dagegen steht das World Wide Web als Quelle für Informationen im Vordergrund.

Die Entwicklungen im Bereich Multimedia haben dazu geführt, daß Daten nicht nur in Form von Texten, sondern auch als Bilder oder Musik vorliegen. Mit Aspekten des Zugriffs auf Multimediadokumente, besonders nicht textueller Dokumente, befaßt sich u.a. [Dunlop und van Rijsbergen, 1993], während es hier um die Gewinnung von Information aus Texten geht. Man unterscheidet drei verschiedene Textsorten: *strukturierte*, *semistrukturierte* und *unstrukturierte* Texte. Der Grad der Strukturierung bezieht sich darauf, wieviel Bedeutung den gegebenen Daten aufgrund ihrer Struktur zugeordnet werden kann. Unstrukturierten Texten wie Zeitungsartikeln kann zunächst keine Bedeutung zugeordnet werden. Zu semistrukturierten Texten zählen z.B. eMails, bei denen die Bedeutung des Eintrages im Absenderfeld bereits bekannt ist. Strukturierte Texte, wie sie z.B. von [Kuikka und Salminen, 1997] verwendet werden, halten sich an ein vorgegebenes Format. Zu jedem Text müssen Angaben in die dafür vorgesehenen Felder eingetragen werden (z.B. Überschrift, Autor, Kapitel, Themenkategorie). Die den Textteilen dadurch zugeordneten Bedeutungen bilden die Grundlage für die Funktionalität des Systems. Das System EMILE II (vgl. [Adriaans et al., 1993]) arbeitet ebenfalls auf strukturierten Texten. Die Texte liegen zwar nicht in einem Format vor, das für jeden Textteil ein bestimmtes Feld vorsieht, aber die Texte weisen nur wenige Variationen im Satzbau auf, und das Thema der Texte ist bekannt. Das System nutzt das Wissen über den Satzbau und das Thema aus, um weitere Informationen aus den gegebenen Texten zu ziehen.

Bei den Texten, die das WWW als Informationsquelle bietet, handelt es sich im oben angeführten Sinn um unstrukturierte Texte. Zwar weisen einige Seiten ähnliche Strukturen auf, und die Frage nach Formatvorgaben steht zur Diskussion, aber die Idee des weltweiten, freien Informationsangebotes und -austauschs steht einer festen Strukturierung der Seiten entgegen.

Zusammenfassend kann also festgehalten werden, daß es um die Aufgabe geht, aus einer Flut von Daten, gegeben als unstrukturierte Texte, interessante Informationen zu gewinnen. Dieser Zielsetzung haben sich die folgenden Bereiche verschrieben:

- Information Retrieval
- Information Filtering und
- Information Extraction.

Die allen drei Bereichen gemeinsame Aufgabenstellung weist auf eine starke Verbindung hin, die sich auch darin zeigt, daß die Bereiche gemeinsame Methoden nutzen. Trotzdem muß zwischen den Bereichen unterschieden werden. Das *Information Retrieval* reagiert auf die Anfrage eines Benutzers. Durch Vergleich der Anfrage mit den Texten der Dokumentenkollektion, die dem IR-System zur Verfügung steht, sucht es mögliche Dokumente, die den Informationsbedarf des Benutzers decken.

Auch ein *Information Filtering*-System stellt dem Benutzer Texte vor. Nach [Belkin und Croft, 1992] decken beide Bereiche allerdings unterschiedliche Informationsbedürfnisse. Während der Benutzer eines IR-Systems einmalig zu einem Thema Information benötigt, weil sein Wissen zu diesem Thema unzureichend ist, versorgt ein IF-System den Benutzer über längere Zeit mit Dokumenten zu einem Thema, für das der Benutzer Interesse bekundet hat. Das IF-System reagiert also nicht auf Anfragen, sondern es erstellt Benutzerprofile, um die Interessen des Benutzers zu beschreiben und mit den Dokumenten zu vergleichen. Dazu benötigt es das Feedback des Benutzers, der die ihm präsentierten Dokumente nach Relevanz gemäß seines Interesses beurteilt.

Information Retrieval und Information Filtering überlassen es dem Benutzer, die eigentlichen Informationen aus den Dokumenten zu ziehen. *Information Extraction* dagegen setzt sich zum Ziel, gesuchte Information aus dem Text zu extrahieren. [Cowie und Lehnert, 1996] vergleichen IR-(IF-)Systeme mit Mähdreschern, die von einem riesigen Feld nutzbares Material in Form von Rohmaterial zurückbringen. IE-Systeme transformieren das Rohmaterial, in dem sie es verfeinern und auf den wichtigen Kern reduzieren.

Neben der Aufgabenstellung haben die drei Bereiche eine weitere Gemeinsamkeit. Seit längerer Zeit werden Techniken aus dem Gebiet *Natural Language Processing* zur Analyse der textuellen Daten eingesetzt. Um die kennzeichnenden Eigenschaften dieser Techniken zu verdeutlichen, führt der folgende Abschnitt 2.1 kurz in dieses Gebiet ein.

Information Retrieval/Filtering und Information Extraction folgen zwei der in der Einleitung vorgestellten Strategien zur Informationsgewinnung: IR und IF treffen eine Vorauswahl aus der Dokumentenkollektion, während IE gezielt Antworten auf Fragen sucht. Mit der dritten Strategie setzt sich ein weiterer Bereich auseinander, die *Clusteranalyse*. Durch die Kategorisierung aller Texte einer Kollektion, d.h. durch Gruppieren von Dokumenten zu gleichen Themen, wird die Datenmenge so strukturiert, daß das Suchen nach Information erleichtert wird. In diesem Bereich sind Techniken aus dem Gebiet NLP nur vereinzelt zu finden. Deshalb werden nach der Einführung in das Gebiet Natural Language Processing in Abschnitt 2.2 Systeme aus den Bereichen Information Retrieval, Information Filtering und Information Extraction vorgestellt, wobei der Schwerpunkt auf dem Einsatz von Techniken zur Verarbeitung natürlicher Sprache liegt. Abschnitt 2.3 stellt danach Systeme zur Textkategorisierung vor.

2.1 Natural Language Processing

Natürliche Sprache wird im Zusammenhang mit Computern von verschiedenen Standpunkten aus betrachtet. Zum Beispiel kann man zwischen gesprochener und geschriebener Sprache differenzieren. Systeme, die gesprochene Sprache verarbeiten, sind mit anderen Problemen konfrontiert als Systeme, die mit geschriebener Sprache arbeiten. Die Informationsquelle WWW bietet Daten in geschriebener Sprache an, weshalb auf die Probleme gesprochener Sprache hier nicht weiter eingegangen werden soll.

Obwohl es hier nicht um natürlichsprachliche Systeme gehen soll, sondern nur um Techniken, die zur Verarbeitung natürlicher Sprache eingesetzt werden, ist hier zum Verständnis des Ausdrucks *Natural Language Processing* die Definition natürlichsprachlicher Systeme nach [Morik, 1995b] angeführt.

Definition: *Natürlichsprachliche Systeme* sind Systeme, die natürliche Sprache analysieren und/oder generieren, wobei sie Wissen über Sprache verwenden.

Auf die Techniken, die natürlichsprachliche Systeme nutzen wird im folgenden mit der Bezeichnung *NLP-Techniken* Bezug genommen. Die entscheidende Eigenschaft, die NLP-

Techniken kennzeichnet, geht aus der Definition hervor: sie verwenden Wissen über Sprache.

Gemäß der Teildisziplinen der Sprachwissenschaft kann das Wissen über Sprache in verschiedene Bereiche unterteilt werden, nach denen im nächsten Abschnitt die NLP-Techniken unterschieden werden. Die Teildisziplinen der Linguistik, die zur Verarbeitung geschriebener Sprache relevant sind, werden nach [Bußmann, 1983] folgendermaßen definiert:

Phonologie Teildisziplin der Sprachwissenschaft, die sich mit den bedeutungsunterscheidenden Sprachlauten (auch: Phonemen), ihren relevanten Eigenschaften, Relationen und Systemen [..] beschäftigt und Methoden zur Analyse von Phonemsystemen bereitstellt.[1]

Morphologie Von Goethe geprägter Terminus zur Bezeichnung der Lehre von Form und Struktur lebender Organismen, der im 19.Jh. als Oberbegriff für Flexion und Wortbildung in die Sprachwissenschaft übernommen wurde.

Syntax System von Regeln, die beschreiben, wie aus einem Inventar von Grundelementen (Morphemen, Wörtern, Satzgliedern) alle wohlgeformten Sätze einer Sprache abgeleitet werden können.

Semantik Bezeichnung von M. Bréal [1897] für die Teildisziplin der Sprachwissenschaft, die sich mit der Analyse und Beschreibung der sog. „wörtlichen" Bedeutung von sprachlichen Ausdrücken beschäftigt.

Pragmatik Aus verschiedenen sprachwiss. philosophischen und sozialwiss. Traditionen hervorgegangene linguistische Teildisziplin, die die Relation zwischen natürlichsprachlichen Ausdrücken und ihren spezifischen Verwendungssituationen untersucht.

Bei der Analyse von Sprache bauen die Teildisziplinen aufeinander auf. Für NLP-Techniken, die Sprache z.B. auf der Ebene der Semantik untersuchen, heißt das, daß die Sprache vorher bereits auf den darunter liegenden Ebenen (Phonologie → Morphologie → Syntax) analysiert werden mußte. Eine Analyse, die bis zur höchsten Ebene, zur Pragmatik, vordringt, und nur mittels NLP-Techniken durchgeführt wird, erfordert also umfangreiches Wissen aus allen Teildisziplinen der Linguistik.

Die Betonung des Kennzeichens, daß NLP-Techniken auf Wissen über Sprache basieren, ist wichtig, um sie von anderen Techniken zu unterscheiden, die in denselben Bereichen für dieselben Aufgaben eingesetzt werden. Ein Beispiel dafür ist Latent Semantic Indexing (LSI). Hierbei handelt es sich um eine Technik, die mittels statistischer Verfahren eine „semantische Struktur" der Wörter innerhalb einer Dokumentenkollektion ermittelt. Nach Reduktion des Wortschatzes aller Dokumente auf wichtige Wörter ergibt sich die semantische Struktur aus dem korrelierten Auftreten der Wörter in den Dokumenten. Die Bezeichnung „semantisch" impliziert nur die Tatsache, daß Wörter eines Dokumentes als Referenz für das Dokument oder für das Thema des Dokumentes behandelt werden ([Deerwester et al., 1990]). Das Verfahren basiert auf der Annahme, daß gemeinsames Auftreten in Dokumenten ein Hinweis darauf ist, daß Wörter die gleiche Bedeutung haben

[1] Die Teildisziplin Phonologie ist hier der Vollständigkeit halber aufgeführt. Kapitel 4 geht noch einmal ausführlicher darauf ein.

oder in einem engen semantischen Zusammenhang stehen. Wissen über Sprache wird nicht verwendet.

Auf ähnliche Weise bildet das oben bereits erwähnte System EMILE II semantische Kategorien: aufgrund der strengen Struktur der Sätze können Wörter, die in demselben syntaktischen Kontext vorkommen, zu einer Kategorie zusammengefaßt werden. Der syntaktische Kontext wird allerdings nicht mit Hilfe von Syntaxregeln bestimmt, sondern durch gleiche Wörter. In dem Beispiel *Derek loves Jean* und *Derek kisses Jean* können *loves* und *kisses* zu einer Kategorie zusammengefaßt werden, weil sie beide innerhalb des gleichen Kontextes (*Derek .. Jean*) verwendet werden können.

Ein drittes Beispiel ist das System NewsWeeder, das in [Lang, 1995] vorgestellt wird. Ob ein Text geeignet ist, die Informationsbedürfnisse eines Benutzers zu decken, wird hier nur anhand statistischer Verfahren getestet.

Neben Techniken, die parallel zu NLP-Techniken angewandt werden können, müssen auch Techniken zur Unterstützung von NLP, die kein Wissen über Sprache verwenden, von NLP-Techniken unterschieden werden. Diese Verfahren dienen z.B. der automatischen Konstruktion eines Thesaurus oder einer Grammatik, die dann von NLP-Verfahren genutzt werden können.

Da der ausschließliche Einsatz von NLP-Techniken auf allen Sprachebenen sehr aufwendig ist, werden meist Techniken, denen kein Sprachwissen zugrundeliegt, mit NLP-Techniken kombiniert. Dabei werden die NLP-Techniken auf den höheren Ebenen eingesetzt, während die darunter liegenden Ebenen mit alternativen Verfahren untersucht werden. Der folgende Abschnitt konzentriert sich bei der Darstellung der Systeme auf die eingesetzten NLP-Techniken. Diese werden danach unterschieden, auf welchen Sprachebenen sie ansetzen, und welche Wissensbereiche über Sprache sie abdecken.

2.2 IR/F- und IE-Systeme unter dem Aspekt NLP

In den Bereichen Information Retrieval, Information Filtering und Information Extraction werden überwiegend englische Dokumente verarbeitet. Das liegt zum einen daran, daß viele Entwickler in diesen Bereichen aus englischsprachigen Ländern kommen. Zum anderen bestehen sehr viele Dokumentenkollektionen aus Texten, die in englischer Sprache verfaßt sind. Für Konferenzen, wie die Text REtrieval Conferences (TREC) und die Message Understanding Conferences (MUC) (vgl. auch [Harman, 1993] und [Sundheim, 1992]), wurden englische Textcorpora zusammengestellt, auf denen verschiedene Systeme im Vergleich getestet wurden. Das erklärt, warum auch Systeme, die in anders sprachigen Ländern entwickelt wurden, zum Teil Englisch verarbeiten. Hinzu kommt, daß Englisch zu den flexionsarmen, einfach strukturierten Sprachen gehört, die einfacher auch ohne Natural Language Processing analysiert werden können. Mit Ausnahme von zwei Systemen, die auf japanische bzw. deutsche Texte angewandt werden, sind alle anderen Systeme, die im folgenden vorgestellt werden, für englische Dokumente konzipiert.

Da wie oben bereits angeführt, die einzelnen Sprachebenen aufeinander aufbauen, beginnt dieser Überblick mit der Ebene der Morphologie und geht dann gemäß der Anordnung in Abschnitt 2.1 jeweils eine Ebene höher. Auf diese Weise soll deutlicher werden, welche Ergebnisse die Analysen auf den einzelnen Ebenen liefern, die dann von den Techniken auf höheren Ebenen aufgegriffen werden.

In [Brill, 1994] geht es noch nicht um ein System, sondern um eine NLP-Technik, deren

Einsatz im Rahmen eines IR-, IF- oder IE-Systems jedoch denkbar ist: *Transformation-based Part of Speech Tagging.* Ziel des Part-of-Speech-Taggings ist es, alle Wörter eines Textes mit einer morphologischen Kategorie zu versehen. Unter morphologischen Kategorien versteht man Wortarten, wie Substantiv, Adjektiv, Verb oder Adverb. Es sind aber auch präzisere Angaben möglich, wie Verb im Präsens oder Substantiv im Singular. Aufgrund der einfachen Struktur der englischen Sprache (aus morphologischer Sicht), die nur wenige Variationen der Wortendungen kennt (Substantive: -s im Plural, Verben: -s in der dritten Person Singular im Präsens, -ing, -ed in der Verlaufsform bzw. im Imperfekt und -ly bei Adverben), werden zum Part-of-Speech-Tagging meist stochastische Verfahren eingesetzt, ohne Sprachwissen zu berücksichtigen. Diese berechnen basierend auf einem Corpus, der manuell mit morphologischen Kategorien versehen wurde, Wahrscheinlichkeiten, mit denen neue Texte analysiert werden.

Von diesen Techniken unterscheidet sich das Transformation-based Part of Speech Tagging durch die Ergänzung durch Regeln, die auf morphologischem Wissen beruhen. Auf diese Weise können u.a. auch unbekannte Wörter mit einem Tag versehen werden. Ein Beispiel für eine solche Regel lautet: „Change tag to **adverb** if the word has suffix **-ly**.“ [Brill, 1994] führt an, daß sich die Erfolgsquote des Part-of-Speech-Taggings mit NLP-Regeln gegenüber ausschließlich stochastischen Verfahren deutlich erhöht.

Das System, das [Strzalkowski, 1995] beschreibt, setzt den Schwerpunkt natürlichsprachlicher Verarbeitung auf der syntaktischen Ebene. Das Vorgehen basiert auf der Annahme, daß die Ergebnisse beim Information Retrieval oder Filtering verbessert werden, wenn neben Wörtern auch zwei- oder mehrelementige Phrasen (z.B. bottle of milk, milk bottle) zur Themenbeschreibung der Dokumente berücksichtigt werden. Auf die bereits morphologisch aufbereiteten Daten wird eine Komponente namens Tagged Text Parser angewandt, der Wissen über die englische Sprache in Form einer Grammatik und dem on-line angebotenen Oxford Advanced Learner's Dictionary nutzt. Ein Parser ist nach [Morik, 1995b] „ein Algorithmus, der einen Satz als Eingabe nimmt und entscheidet, ob er bezüglich einer gegebenen Grammatik wohlgeformt ist, und ihm - falls ja - eine Strukturbeschreibung zuordnet“ (S.14). Dieser Strukturbeschreibung, die jedem Wort eine Funktion im Satz zuweist (Subjekt, Prädikat, Objekt etc.) können dann die Phrasen entnommen werden. Auch [Strzalkowski, 1995] merkt an, daß statistische Verfahren eine zu hohe Fehlerrate aufweisen, um Phrasen präzise zu ermitteln.

Das System BREVIDOC [Miike et al., 1994] verwendet auf allen Ebenen linguistisches Wissen und ist eins von den beiden Ausnahmesystemen, denn es analysiert japanische Dokumente. Obwohl hier auch die morphologische und die syntaktische Analysekomponente interessant sind, wird nur die Ebene der Semantik aufgezeigt.

Das System verfolgt die Zielsetzung, die Präzision beim Information Retrieval zu erhöhen und dem Benutzer das Suchen der Information in den ermittelten Dokumenten zu erleichtern. Zu diesem Zweck werden im Anschluß an die syntaktische Analyse Phrasen wie „Mit anderen Worten...“, „Im Gegensatz...“ oder „Der Grund ist...“ gesucht, die Rückschlüsse auf die Struktur des Dokumentes und auf die semantische Rolle eines Satzes zulassen. Ein Satz, der von einer Phrase wie „Der Grund ist...“ eingeleitet wird, übernimmt die semantische Rolle einer Begründung. Da die Reihenfolge von Begründungen, Beispielen, Wiederholungen u.ä. den Aufbau eines Dokumentes bestimmen, kann aufgrund der semantischen Rollen der Sätze auf die Struktur des Dokumentes geschlossen werden. Außerdem läßt sich aus der semantischen Rolle eines Satzes seine Bedeutung für das gesamte Dokument ableiten. Ein Beispiel trägt zur Bestimmung des Dokumentthemas weniger bei als ein zusammenfassender Satz. Die Zuordnung von semantischen Rollen zu Sätzen ermöglicht die Erfüllung der gesteckten Ziele: Enthält ein Satz ein Wort, das ebenfalls in der Anfrage enthalten ist, dann beeinflußt die semantische Rolle des Satzes die Bewertung des Wortes bezüglich der Anfrage. Ist das Anfragewort nur in einem Beispielsatz enthalten, kann nicht mit Sicherheit festgestellt werden, ob das Dokument in Bezug auf die Anfrage relevant ist. Damit wird die Präzision des Retrievalsystems erhöht. Die Analyse der Struktur dient der Zusammenfassung des Dokumentes. Wenn gemäß einer Anfrage entsprechende Dokumente ermittelt wurden, stellt BREVIDOC dem Benutzer die kürzeren Zusammenfassungen vor, um ihm die weitere Suche zu erleichtern.

BREVIDOC verwendet also nicht nur morphologisches und syntaktisches Wissen, sondern auch umfangreiches Wissen über die Semantik rhetorischer Phrasen. Neben Lexikon und Grammatik greift es auf drei weitere Wissensbasen zu, die aber alle drei nur linguistisches Wissen enthalten. Dadurch ist das System domänenunabhängig einsetzbar.

Das von [Wendlandt und Driscoll, 1991] beschriebene System wendet ebenfalls Konzepte aus der Semantik an: thematische Relationen und semantische Merkmale. Zur Definition thematischer Relationen schreibt [Bußmann, 1983]: „Von J.S. Gruber[1967] entdeckte und von R.S. Jackendoff [1972] genauer erarbeitete, kasusähnliche semantische Relationen, die die Beschreibung verschiedener gramm. Probleme vereinfachen. In dem Satz *Caroline leiht sich zu Hause das Zauberbuch von Philip* werden den Nominalphrasen folgende thematische Relationen zugeschrieben: *Caroline = Agent* und *Goal, zu Hause = Location, das Zauberbuch = Theme* und *Philip = Source*“ (S.542). Weitere Relationen, die [Wendlandt und Driscoll, 1991] anführen, lauten: *Duration, Time* und *Motion*.

Während die thematischen Relationen, Beziehungen zwischen Objekten beschreiben, konkretisieren semantische Merkmale Eigenschaften der Objekte. Zu den semantischen Merkmalen gehören Angaben in Bezug auf Größe, Farbe u.ä. Bei [Wendlandt und Driscoll, 1991] geht es aber nicht um die Angabe semantischer Merkmale (wie *blau* oder *rot*), sondern nur um die Zuweisung von Merkmalkategorien (wie *Farbe* oder *Größe*). Untersucht wird, welche Wörter des Textes auf welche Relation und welche Merkmalkategorie hinweisen. So deuten z.B. Wörter wie *dimension, feet, diameter* und *long* Merkmale der Kategorie Größe an. Die Frage *How long does the payload crew go through training before launch?* enthält mehrere Wörter, die thematische Relationen anzeigen: *how long = Duration/Time, through = Duration/Motion, before = Location/Time*.

Beim Information Retrieval/Filtering werden die thematischen Relationen und semantischen Merkmalkategorien genutzt, in dem sowohl die Wörter der Anfrage und der Dokumente verglichen werden, als auch die implizierten Relationen und Merkmalkategorien. Für diese Retrievalstrategie ist ein Lexikon erforderlich, das den entsprechenden Wörtern thematische Relationen oder semantische Merkmalkategorien zuordnet. Nur solange die Texte aus einer speziellen Domäne stammen und das Lexikon nicht den gesamten Wortschatz einer Sprache umfaßt, ist die manuelle Zuordnung möglich. Zwar ist der Wortschatz einer Sprache nicht unbeschränkt, so daß bei entsprechendem Aufwand auch der ganze Wortschatz bearbeitet werden könnte, problematisch ist aber eine vollständige Festlegung aller Merkmalkategorien, weil die meisten Kategorien von einem Sachbereich abhängig sind.

[Mauldin, 1991] stellt das System FERRET als *Conceptual Information Retrieval System* vor. Es stellt einen Versuch dar, Texte in Frames zu überführen und diese dann mit einer geeigneten Darstellung der Anfrage zu vergleichen. Problematisch erwies sich allerdings die Evaluierung des Systems, weil das System keine Komponente zur Umsetzung der Anfrage in eine geeignete Darstellung bereitstellt. Damit bleibt die Frage offen, ob eine Darstellung der Dokumente als Frames die Retrievalperformanz erhöhen kann, besonders weil durch Verwendung manuell anzulegender Frames die Domäne eingeschränkt werden muß.

Die Anwendung von Frames oder Skripten ist typisch für Systeme aus dem Bereich Information Extraction, deren Ziel es ist, Fragen des Benutzers direkt zu beantworten. Zwei typische Systeme sollen hier vorgestellt werden: das System FRUMP [DeJong, 1982] und das System SCISOR [Jacobs und Rau, 1988], [Jacobs und Rau, 1990]. Beide Systeme beziehen Weltwissen in die natürlichsprachliche Verarbeitung ein. [DeJong, 1982] bezeichnet dieses Weltwissen als „pragmatisches Wissen des Systems" (S.149). Ob es sich aber tatsächlich um Wissen auf der Ebene der Pragmatik handelt, ist fraglich. Aus der oben angeführten Definition geht hervor, daß die Pragmatik die Beziehung natürlichsprachlicher Ausdrücke zu ihren spezifischen Verwendungssituationen untersucht. Das hieße, daß pragmatisches Wissen es ermöglicht, gleichen Wörtern in verschiedenen Verwendungssituationen verschiedene Bedeutung zuzuweisen. Das von beiden Systemen eingesetzte Weltwissen ist aber nur darauf ausgelegt, die Analyse von Texten aus einer festgelegten Informationsquelle zu bestimmten Themen zu unterstützen. Damit bleiben auch diese Systeme auf der Ebene der Semantik.

FRUMP und SCISOR weisen einige Gemeinsamkeiten auf. Beide trennen linguistisches Wissen von dem Wissen über die Welt. Die linguistische Analyse geht über alle Ebenen bis zur Semantik und endet mit der Zuordnung thematischer Relationen, die bereits vorgestellt wurden. Für SCISOR beschreibt [Jacobs und Rau, 1990] dies als „linguistische bottom-up Analyse". Die Verwendung des Weltwissens dagegen erfolgt top-down. Wenn bestimmte Wörter im Text auftreten, werden Skripten aktiviert, die dann gemäß anderer Wörter, die im Kontext des auslösenden Wortes stehen, ausgefüllt werden. Durch Abgleich der Skripten mit den Ergebnissen der linguistischen Analyse können Fehler der einen oder anderen Methode verbessert werden oder fehlende Information ergänzt werden. Die Sätze eines Textes werden auf diese Weise analysiert, und die zu einem Text ermittelten Skripten werden, wenn möglich, zueinander in Beziehung gesetzt und in die Wissensbasis aufgenommen. Benutzerfragen werden auf dieselbe Weise analysiert. Das entstandene Skript wird mit der Wissensbasis abgeglichen, um eine Antwort zu extrahieren.

Beide Systeme verarbeiten on-line angebotene Pressemeldungen. SCISOR ist auf Meldungen aus der Finanzwelt spezialisiert. FRUMP dagegen sieht Skripten für sechzig Ereignistypen vor, über die in Pressemeldungen berichtet wird (z.B. Erdbeben oder Streiks). Zwar argumentiert [DeJong, 1982], daß er nicht auf eine bestimmte Domäne beschränkt ist, weil Meldungen aus vielen Bereichen verarbeitet werden können, aber auch seine Wissensbasis ist auf diese sechzig Ereignistypen begrenzt. Wenn SCISOR sechzig verschiedene Ereignisse aus der Finanzwelt unterscheiden kann, ist es trotz Beschränkung der Domäne genauso leistungsfähig. Die Trennung der beiden Wissens- und Analysebereiche (Sprache und Welt) ermöglicht es allerdings, die Systeme auch auf anderen Gebieten einzusetzen. Das Weltwissen muß dann an das jeweilige Gebiet angepaßt werden, was sehr aufwendig ist.

Das letzte hier vorzustellende System namens TOPIC [Hahn und Reimer, 1986] basiert auf Wissen über die deutsche Sprache. Genauso wie BREVIDOC ist es konzipiert, Texte zusammenzufassen, allerdings nicht um das Retrieval zu erleichtern, weshalb es den

Systemen aus dem Bereich Information Extraction zugeordnet ist. Analog zu BREVIDOC analysiert TOPIC die Struktur eines Textes und kombiniert dann die wichtigsten Aussagesätze. Dabei wird der Text wie bei SCISOR und FRUMP als Kombination von Frames dargestellt. Interessant an diesem System ist, daß zur Analyse des gesamten Textes nicht nur alle einzelnen Satzanalysen zusammengetragen werden, sondern der Text satzübergreifend analysiert wird.

Dieser Ausschnitt existierender Systeme sollte die Möglichkeiten aufzeigen, auf welchen Ebenen und in welchem Umfang NLP-Techniken bei der Analyse von Textdaten zum Einsatz kommen können. Der folgende Abschnitt präsentiert Systeme, die die Informationsgewinnung durch Kategorisierung von Textdaten unterstützen.

2.3 Textkategorisierung

Zur Kategorisierung der Texte einer Dokumentenkollektion wird ein Verfahren angewandt, das auch auf anderen Gebieten wie der Wissensentdeckung in Datenbanken eingesetzt wird: die *Clusteranalyse*. Ziel der Clusteranalyse ist es, die Struktur, die einer gegebenen, ungeordneten Menge von Objekten (hier Dokumenten) zugrundeliegt, ausfindig zu machen. D.h. unter der Voraussetzung, daß diese Struktur tatsächlich existiert, werden die Objekte bezüglich ihrer Ähnlichkeit gruppiert. Für Texte heißt das z.B., daß Texte mit einem gemeinsamen Thema nah beieinander angeordnet werden und zu Kategorien zusammengefaßt werden können, während sie gleichzeitig von Texten mit anderen Themen getrennt werden. Abschnitt 4.3 geht näher auf die Clusteranalyse ein und stellt verschiedene Methoden dar. Im folgenden werden vier Systeme vorgestellt, die verdeutlichen, wie die Clusteranalyse zur Unterstützung der Informationsgewinnung eingesetzt werden kann.

Das System Scatter/Gather [Cutting et al., 1993], [Hearst et al., 1995] bietet verschiedene Modi an. Zum einen kann die gesamte Dokumentenkollektion strukturiert werden, oder es wird auf die Ergebnisse eines Retrievals angesetzt, um die auf Anfrage ermittelten Dokumente so zu Gruppen zusammenzufassen, daß der Benutzer schneller auf Dokumente stößt, die die gesuchte Information enthalten. Dazu werden dem Benutzer Beschreibungen der Cluster präsentiert. Diese Beschreibungen bestehen aus den Titeln einiger „typischer" Dokumente eines Clusters und aus einer Liste der Wörter, die das Cluster charakterisieren (vgl. Abschnitt 4.2). Scatter/Gather kann auch wiederholt angewandt werden, in dem zunächst aus der gesamten Kollektion grobe Cluster ausgewählt werden, die dann noch einmal zu feineren Clustern gruppiert werden. Durch die Wiederholung dieses Vorgehens werden die Cluster kleiner und die Beschreibungen präziser.

[Sahami et al., 1996] wenden das Multiple Cause Mixture Model (vgl. [Saund, 1995]), ein Modell aus dem Bereich Neuronale Netze, zur Textkategorisierung an. Das Modell zeichnet sich dadurch aus, daß Dokumente auch mehreren Kategorien (Themen) zugeordnet werden können. Jedes Thema wird durch eine Anzahl Wörter beschrieben, wobei die Wortmengen nicht disjunkt sind. Enthält ein Dokument die Schnittmenge der Wortmengen zweier Themen, werden dem Dokument beide Themen zugewiesen. Ursprünglich war das Modell nur zur Klassifikation von Dokumenten vorgesehen. Dabei zeigte sich, daß es auch zur Clusteranalyse eingesetzt werden kann (zur Differenzierung Clusteranalyse $\leftrightarrow$ Klassifikation siehe Abschnitt 4.3), wobei der Schwerpunkt allerdings weiter auf der Klassifikation liegt.

Das in [Joachims et al., 1995] vorgestellte System WebWatcher ist zur Unterstützung des Browsings im World Wide Web konzipiert. Das System beobachtet das Verhalten eines Benutzers während des Browsings und lernt daraus. Ist dem System bekannt, für welche Seiten ein Benutzer Interesse bekundet hat, kann es einem anderen Benutzer, der sich für eine Teilmenge dieser Seiten interessiert, die anderen Seiten ebenfalls empfehlen. Diese Strategie geht auf die Idee des Citation Clusterings zurück, das in Abschnitt 4.3 vorgestellt wird.

Die vorhergehenden Systeme setzen keine NLP-Techniken ein. Das System NLDB [Rau und Jacobs, 1991] dagegen nutzt semantisches Wissen, um die Dokumentenkollektion zu kategorisieren. Um festzustellen, über welches Thema ein Dokument berichtet, wird eine Methode namens Pattern Matching aktiviert. Anstatt Schlüsselwörter eines Textes zu berücksichtigen, wird untersucht, ob ein Text bestimmte Wortfolgen (= Patterns) enthält, die für ein Thema charakteristisch sind. Beispielsweise weist die Wortfolge $< company > \cdots < acquire > \cdots < company >$ auf das Thema „Firmenfusion" hin. Dem Pattern Matching geht eine morphologische Analyse der Wörter voraus, damit Sätze, die Wörter wie *acquisition* statt *acquire* enthalten, ebenfalls entdeckt werden können. Die syntaktische Analyse ermöglicht es, Satzkonstruktionen auszuschließen, die durch Einschübe wie Relativsätze die gesuchten Wortfolgen erzeugen, ohne daß das entsprechende Thema berührt wird (Bsp.: *... GE, whose acquisition of RCA led to increased earnings, ...*). Gemäß der Differenzierung zwischen Clusteranalyse und Klassifikation, die in Abschnitt 4.3 vorgenommen wird, ist NLDB kein echtes Clusterverfahren, denn durch Vorgabe der Wortfolgen und der zugehörigen Themen wird der Gruppierung der Dokumente bereits vorgegriffen. Trotzdem ist das System wegen der verwendeten NLP-Techniken an dieser Stelle aufgeführt.

3 Besonderheiten der Domäne

Die Domäne, mit der sich diese Arbeit auseinandersetzt, heißt *deutsche Tageszeitungen*. Schon diese Bezeichnung impliziert eine Zweiteilung. Zum einen geht es bei der zu verarbeitenden Sprache um *Deutsch* statt um Englisch wie in zahlreichen anderen Systemen, zum anderen hat der Bereich *Tageszeitung* beachtenswerte Merkmale. Deshalb ist auch dieses Kapitel zweigeteilt: Der erste Abschnitt befaßt sich mit dem Gebiet Zeitungswesen, dessen Sprache einen eigenen Forschungsbereich in der Linguistik einnimmt. Der zweite Abschnitt geht auf einige Besonderheiten der deutschen Sprache im Vergleich zur englischen ein und diskutiert, wie Natural Language Processing im Rahmen dieser Arbeit eingesetzt werden kann.

3.1 Pressesprache

Presse- oder Zeitungssprache wird in der Linguistik unter verschiedenen Gesichtspunkten betrachtet. Hier sollen die Ergebnisse der Untersuchungen bezüglich Syntax und Wortschatz im Vordergrund stehen, weil sich aus ihnen Kennzeichen der Pressesprache ableiten lassen, die sich auf den Entwurf des Systems auswirken. Syntax und Wortschatz in Zeitungen werden unter drei (groben) Aspekten untersucht (vgl. [Lüger, 1995]):

1. Pressesprache als Indiz für Tendenzen der Gegenwartssprache,

2. Pressesprache als spezifischer Sprachgebrauch im Medium Presse,

3. Pressesprache als Sprachgebrauch eines bestimmten Publikationsorgans.

Der erste Aspekt deutet an, daß sich die Eigenschaften der Pressesprache auf die Gegenwartssprache verallgemeinern und eventuell auf andere Bereiche übertragen lassen. Der zweite Aspekt bestätigt Vermutungen, daß sich die Pressesprache durch einen bestimmten Stil auszeichnet, der im weiteren genutzt werden kann. Der letzte Aspekt weist darauf hin, daß dieser Bereich differenziert untersucht werden muß. Man muß zwischen Zeitungen und Zeitschriften, regionalen und überregionalen und zwischen Tages- und Wochenzeitungen unterscheiden. Hinzu kommt, welcher Leserkreis angesprochen wird. Die im folgenden dargestellten, allgemeinen Kennzeichen der Pressesprache lassen sich weitgehend auf alle Teilgebiete übertragen und gelten insbesondere für Tageszeitungen, bei denen viele Untersuchungen ansetzen.

Die Entwicklung des Zeitungsstils wird auf die Produktionsbedingungen zurückgeführt, unter denen Pressetexte entstehen: verschiedene Agenturmeldungen und andere Quellen müssen zusammengefaßt werden, wobei möglichst viel Information auf engem Raum gegeben werden soll (vgl. [Lüger, 1995]). Für die Syntax in Zeitungsartikeln ergibt sich damit:

- die Tendenz zur Verkürzung der Satzlänge,
- der Rückgang von Satzgefügen zugunsten von Einfachsätzen,
- der Übergang vom Verbalstil zum Nominalstil.

Die Tendenz zur Verkürzung der Satzlänge schreibt [Raith, 1988] der Forderung nach Lesbarkeit journalistischer Texte zu. Die (berechtigte) Kritik, daß lange Schachtelsätze das Verständnis erschweren, führt zu Darstellungen in möglichst kurzen Sätzen. Statistische Erhebungen belegen nach [Eggers, 1973], daß die Verwendung von Satzgefügen zurückgegangen ist, während die Verwendung von Einfachsätzen zugenommen hat[2]. Der Übergang zu Einfachsätzen, in denen genauso viel Information untergebracht werden muß wie in Satzgefügen, wird begleitet von der Tendenz zum Nominalstil in Pressetexten. [Bußmann, 1983] (S.350) definiert:

> **Nominalstil**. Durch Vorherrschen von Substantiven gegenüber (finiten) Verben gekennzeichneter Sprachstil, der in neueren Stilistiken als Hauptkennzeichen der Gegenwartssprache beschrieben und häufig als „Papierstil“, „Kanzleideutsch“, „Hauptwörterseuche“ kritisiert wird.

Diskussionen über die Berechtigung dieser Kritik werden von vielen Autoren wie z.B. [Seiffert, 1977] und [Raith, 1988] geführt. In der Praxis setzt sich der Nominalstil trotzdem durch. Zu den Merkmalen des Nominalstils zählen attributive Erweiterungen (*die Zustimmung der Mitglieder*), präpositionale Fügungen (*unter Verzicht, durch wachsendes Interesse*) und Verbaufspaltungen (*zur Durchführung bringen, eine Mitteilung machen, einen Versuch unternehmen* statt *durchführen, mitteilen, versuchen*). Zu den Verbaufspaltungen bemerkt [Lüger, 1995] (S.26): „die eigentliche Bedeutung liegt jeweils auf dem

[2][Lüger, 1995] definiert: *Einfachsätze* bestehen aus nur einem Hauptsatz, ohne Nebensatz oder satzwertigen Infinitiv. [..] *Satzgefüge* weisen außer dem Hauptsatz wenigstens einen Nebensatz oder satzwertigen Infinitiv auf.“ (S.24)

nominalen Teil des Ausdrucks, die Bedeutung der Verben *bringen, machen, unternehmen* usw. ist abgeschwächt.“

Eine weitere Eigenschaft, die die Pressesprache kennzeichnet, ist die Syntax der Überschriften. Diese werden meist verkürzt, um Wiederholungen in Überschrift und Text zu verhindern. Zur Verkürzung werden zwei Strategien angewandt: die *Ersparung* und die *Auslassung*. Um eine Ersparung handelt es sich dann, wenn die Überschrift aus einem unvollständigen Satz besteht, dessen fehlende Teile vom Leser trotzdem problemlos mitverstanden werden („*KSC in letzter Sekunde für Masters qualifiziert*“ [SVZ, 23.01.97], „*Peru im Schatten der Geiselnahme*“ [PNP, 23.01.97]). Auslassungen (oder Ellipsen) dagegen sind so stark verkürzt, daß der Leser Fehlendes nicht mehr ergänzen kann und auch nicht auf den Inhalt des Textes schließen kann (*Was fehlt* , *Vom Schlechten des Guten* [beides TAZ, 23.01.97]). [Schneider und Esslinger, 1993] warnt davor, die Überschrift zu stark zu verkürzen: die Unterschlagung von Subjekt oder Dativobjekt verletze das Sprachgefühl und sei nur schwer erträglich. Hier wird deutlich, daß Verkürzungen meist nur Wortarten wie Verben, Präpositionen, oder Artikel betreffen sollten, aber keine Substantive.

Neben die typischen, syntaktischen Merkmale der Pressesprache treten die Merkmale bezüglich des Wortschatzes: „Lexikalisch auffallend ist die Verwendung immer neuer (aktueller) Bezeichnungen, fachsprachlicher Ausdrücke und Fremdwörter sowie die (Wort)Bildung von Ad-hoc-Zusammensetzungen“ ([Lewandowski, 1994], S.1280). Die „Verwendung neuer Bezeichnungen, die in den gängigen Wörterbüchern (noch) nicht verzeichnet sind“, nennt [Lüger, 1995] (S.30) das „auffallendste Kennzeichen“ und er zitiert [Eggers, 1973] (S.99), der aufzeigt, daß die relative Häufigkeit dieser Wortschöpfungen über das „wechselnde Zeitinteresse“ Aufschluß gibt: man könnte „aus Tageszeitungen von Jahr zu Jahr, aber auch für jeden beliebigen einzelnen Tag ermitteln, wie sich die Interessen der öffentlichen Diskussion innerhalb eines bestimmten Zeitraums verteilen, oder was zu bestimmter Zeit das ‘Tagesgespräch’ war. Am Emporschnellen der relativen Häufigkeit des einschlägigen Vokabulars ließe sich das unschwer ablesen.“ In der Regel handelt es sich sowohl bei den Wortschöpfungen als auch bei den fachsprachlichen Ausdrücken und Fremdwörtern um Substantive. Ad-hoc-Zusammensetzungen (auch Augenblickskomposita) sind eine Erscheinung, die ebenfalls dem Zwang zur Kürze zuzuschreiben ist. Durch diese zusammengesetzten Ausdrücke können mehrere Informationseinheiten komprimiert wiedergegeben werden (*die Vereinbarung zum Aufschub der Wahl* → *die Aufschub-Vereinbarung*). Augenblickskomposita erlauben es also, verkürzt auf einen bereits erwähnten Sachverhalt einzugehen, ohne ihn noch einmal wörtlich zu wiederholen.

Die angeführten syntaktischen und lexikalischen Merkmale der Pressesprache, die Verkürzung der Satzlänge, der überwiegende Gebrauch von Einfachsätzen und daraus folgend das Vorherrschen des Nominalstils, die besondere Syntax der Überschriften und die Verwendung neuer Bezeichnungen, fachsprachlicher Ausdrücke, Fremdwörter und Ad-hoc-Zusammensetzungen, weisen auf die besondere Bedeutung von Substantiven in Zeitungsartikeln hin: der Nominalstil, der auch die Aufspaltung von Verben in bedeutungstragende Substantive und abgeschwächte Verben mit sich bringt, führt dazu, daß Sätze zu einem großen Teil aus Substantiven bestehen, die durch bedeutungsschwächere Verben ergänzt werden. Die lexikalischen Merkmale betreffen in den meisten Fällen Substantive, was deren Wichtigkeit ebenfalls unterstreicht.

Die Verwendung einer weiteren Wortart ist genauso umstritten und wird ähnlich diskutiert wie der Nominalstil: die Verwendung von Adjektiven. Viele Autoren wie [Schneider, 1984] bemängeln besonders Tautologien, die durch gedankenloses Anfügen von Adjektiven an Substantive entstehen, und den Bezug von Adjektiven zu falschen Substantiven: Adjektive „produzieren Tautologien (weiße Schimmel, aber 'schwere Verwüstungen' sind um nichts besser, denn wer hätte je *leichte* Verwüstungen gesehen?) [..] Den äußersten Unfug richten die Adjektive dort an, wo sie die Logik auf den Kopf stellen, weil sie aufs falsche Substantiv bezogen werden. [..] zweimal in der Woche lassen wir uns *atlantische Tiefausläufer* bieten, obwohl hier doch nicht irgendein Tief atlantisch ausläuft, sondern ein Atlantiktief seine Ausläufer schickt“ (S.37-40). [Raith, 1988] setzt dagegen, daß auch „leere Worte“ wie „*drastische* Preissteigerung“ etwas aussagen können und nicht unbedingt ein Zeichen für die Gedankenlosigkeit des Verfassers sein müssen. Er empfiehlt einen sorgsamen Umgang mit Adjektiven, wobei nur überflüssige Adjektive weggelassen und bedeutungstragende effektiv eingesetzt werden.

Die Diskussion um die Adjektive zeigt, daß sie oft (wenn auch gedankenlos oder falsch) verwendet werden. Außerdem führen Fürsprecher an, daß nicht alle Adjektive gestrichen werden können, weil sie elegant weitere Information liefern können (z.B. Adjektive wie *französisch, europäisch, sportlich* etc.).

Der nächste Abschnitt geht auf den Entwurf des vorliegenden Systems ein, wobei die hier beschriebenen Merkmale der Pressesprache und weitere Eigenschaften der deutschen Sprache besonders berücksichtigt werden.

3.2 Deutsche Sprache – schwere Sprache

In Abschnitt 2.1 und 2.2 wurden die verschiedenen Teildisziplinen der Linguistik vorgestellt, an denen NLP-Techniken ansetzen können, und deren Wissen über Sprache sie nutzen können. Aufgrund der Unterschiede zwischen der englischen und der deutschen Sprache können die in den angeführten Systemen verwendeten NLP-Techniken nicht ohne weiteres in deutschen Systemen eingesetzt werden. Zu den Unterschieden gehören z.B. der Flexionsreichtum des Deutschen im Vergleich zum Englischen und die Variationsmöglichkeiten, die im Deutschen in der Satzstellung erlaubt sind. Englische Sätze folgen im wesentlichen dem Aufbau Subjekt–Prädikat–Objekt, während diese Satzstellung im Deutschen meist zur Betonung bestimmter Satzteile variiert werden darf. Diese Unterschiede deuten bereits an, daß ein System, das zur Erfassung der deutschen Sprache ohne Einschränkung der Domäne konzipiert ist, auf NLP-Techniken angewiesen ist, weil statistische Methoden zur genauen morphologischen oder syntaktischen Analyse nicht ausreichen.

Auf der semantischen Ebene bedürfen beide Sprachen eines umfangreichen zusätzlichen Wissens und damit einhergehend einer Einschränkung der Domäne. Zwar handelt es sich bei der Verarbeitung von Zeitungsartikeln auch um eine Einschränkung auf den Bereich Zeitungswesen, aber die Ereignisse, über die die Artikel berichten, sind nicht auf bestimmte Typen oder Themen beschränkt (wie bei FRUMP oder SCISOR). Alle Artikel werden bearbeitet. Auch die Konzentration auf die Eigenschaften der Pressesprache bedeutet keine Einschränkung. So schreibt [Lewandowski, 1994] zum Nominalstil: „Syntaktische Strategie in der Gegenwartssprache (Wissenschaftssprache, Publizistik, Verwaltung)“ (S.748) und überträgt damit dieses herausragende Merkmal der Pressesprache auch auf andere Bereiche. [Lüger, 1995] stellt fest, daß „Entwicklungen und charakteristische Merkmale der deutschen Gegenwartssprache“ am Beispiel der Pressesprache dargestellt werden, und daß die Ergebnisse „von spezielleren Untersuchungen [..] weitgehend bestätigt und in einzelnen

Punkten ergänzt“ werden (S. 22f).

Im Rahmen dieser Arbeit wird das Wissen über Sprache auf den Bereich Morphologie beschränkt. Zwar ist der Einsatz einer Syntaxanalyse in diesem Bereich ebenfalls denkbar, um Ideen wie die Extraktion von Phrasen für die deutsche Sprache umzusetzen, aber ein System, das die Syntaxanalyse übernimmt, konnte nicht ausfindig gemacht werden. Der Einsatz einer semantischen Komponente hätte eine Einschränkung der Anwendungsdomäne bedeutet. Deshalb wurde auf die Nutzung semantischen Wissens verzichtet. Außerdem reagieren syntaktische und semantische Komponenten oft empfindlich auf unbekannte Wörter. Pressesprache zeichnet sich aber durch das häufige Auftreten von Wortschöpfungen (z.B. Augenblickskomposita) aus. Diese müssen bei der Analyse unbedingt berücksichtigt werden, weil gerade sie das ‘Tagesgespräch’ charakterisieren (vgl. 3.1).

Der Aufbau des gesamten Systems, besonders auch der Clusteranalyse wird in Kapitel 4 dargestellt. Der Rest dieses Abschnitts stellt die Idee vor, wie die morphologische Analyse unter Berücksichtigung der Merkmale der Pressesprache zur Erschließung der Artikelthemen eingesetzt werden kann.

Mit den Ausführungen des vorhergehenden Abschnitts soll dem Verb nicht seine Bedeutung für die deutsche Sprache genommen werden, denn „das Verb repräsentiert nach z.T. älterer, vor allem aber neuerer Auffassung die wichtigste Wortart. Das Verb verfügt über den größten Formen- und Funktionsreichtum; es ist das satzbildende Wort, der strukturelle Kern, das eigentliche Zentrum des Satzes. Aus psycholinguistischer Sicht bestimmt das Verb die Rollen, die die Nominalphrasen im Satz spielen, und ihre semantischen Merkmale bzw. allgemeinen Bedeutungen“ ([Lewandowski, 1994], S.1221). Das Verb bestimmt also die Struktur des Satzes. Im Rahmen der hier gestellten Aufgabe wird aus den genannten Gründen zur Erschließung eines Textthemas die Struktur vernachlässigt. Stattdessen wird der Inhalt eines Dokumentes unter der Annahme untersucht, daß aufgrund der Eigenschaften der Pressesprache, die Information überwiegend in Nominalphrasen formuliert, Substantive (dazu werden hier auch Eigennamen gezählt) und Adjektive das Thema eines Textes repräsentieren. Das folgende Beispiel soll dies verdeutlichen:

> **„Die Tür ist geöffnet“**
>
> Eine der letzten Männerbastionen im Bereich der internationalen Kultur fällt. Die Wiener Philharmoniker werden künftig Frauen aufnehmen. Das bestätigte Orchestervorstand Werner Resel gestern im österreichischen Fernsehen. „Die Tür ist geöffnet“, sagte Resel. Das Orchester, das als eines der besten der Welt gilt, wolle sich „dem Trend der Zeit nicht verschließen“. Die Vorstandsentscheidung, die nach einem Gespräch mit dem Kunstministerium zustande gekommen war, muß allerdings noch von der Vollversammlung des Orchesters bestätigt werden. [..] (Passauer Neue Presse, 24.01.97)

Dieser Text gibt einen Ausschnitt aus einem Artikel der Passauer Neuen Presse vom 24.01.1997 wieder. Die beiden folgenden Texte beruhen auf diesem Ausschnitt. Allerdings fehlen im linken Text die Adjektive, Substantive und Eigennamen, während der rechte Text nur diese drei Wortarten enthält.

„Die .. ist geöffnet“

Eine der .. im .. der .. fällt. Die .. werden künftig .. aufnehmen. Das bestätigte .. gestern im .. „Die .. ist geöffnet“, sagte .. Das .., das als eines der .. der .. gilt, wolle sich „dem .. der .. nicht verschließen“. Die .., die nach einem .. mit dem .. zustande gekommen war, muß allerdings noch von der .. des .. bestätigt werden.

„.. Tür ..“

.. letzten Männerbastionen .. Bereich .. internationalen Kultur .. Wiener Philharmoniker .. Frauen .. Orchestervorstand Werner Resel .. österreichischem Fernsehen .. Tür .. Resel .. Orchester .. besten .. Welt .. Trend .. Zeit .. Vorstandsentscheidung .. Gespräch .. Kunstministerium .. Vollversammlung .. Orchesters ..

Der Leser des linken Textes hat keine Möglichkeit ohne Kenntnis des originalen Ausschnitts auf das Thema des Textes zu schließen. Der Leser des rechten Textes dagegen entwickelt eine Vorstellung vom Textthema. Damit liegt es nahe, die morphologische Analyse zu nutzen, um bedeutungstragende Wortarten aus einem Text zu extrahieren, und somit das Thema eines Textes zu beschreiben.

4 Das System AKAT

Die Ausführungen in Kapitel 2 haben Möglichkeiten aufgezeigt, wie Techniken aus dem Gebiet Natural Language Processing zur Unterstützung der Informationsgewinnung eingesetzt werden können. Kapitel 3 hat verdeutlicht, welche Besonderheiten die deutsche Sprache und die Domäne Tageszeitungen aufweisen, die ausgenutzt werden können. Zusammen mit der Aufgabenstellung, dem Zusammenstellen von Artikeln mit einem gemeinsamen Thema, ergibt sich der folgende Aufbau des Systems AKAT, dessen Arbeitsschritte Abbildung 1 zeigt.

Abbildung 1: Ablaufdiagramm

Zunächst werden die Zeitungsartikel, die als Volltexte vorliegen, einer *morphologischen Analyse* unterzogen. Diese Aufgabe übernimmt das System Gertwol, das in Abschnitt 4.1 vorgestellt wird.

Die Ergebnisse der morphologischen Analyse werden genutzt, um die Texte im Rahmen der *automatischen Indexierung* zu filtern. In diesem Arbeitsschritt werden die gefilterten Texte auch in eine Repräsentation überführt, die es möglich macht, die Texte miteinander zu vergleichen. Abschnitt 4.2 beschreibt die Verfahren zur automatischen Indexierung.

Zur Kategorisierung der Texte wird eine *Clusteranalyse* durchgeführt. Abschnitt 4.3 zeigt verschiedene Verfahren zur Clusteranalyse auf und stellt das hier angewendete Verfahren vor.

4.1 Morphologische Analyse

Die Überlegungen in Kapitel 3 legen nahe, die Clusteranalyse durch Informationen zu unterstützen, die auf Wissen über die Sprache beruhen. Dazu können verschiedene NLP-Techniken eingesetzt werden. Die Systeme, die in Abschnitt 2.2 vorgestellt wurden, lassen sich bezüglich der Strategie zur Beschreibung der Dokumente in zwei Kategorien einteilen: Systeme wie FRUMP oder SCISOR überführen die Texte in komplexe Strukturen, während andere Systeme auf verschiedene Weise bedeutungstragende Wörter oder Phrasen zur Beschreibung der Dokumente extrahieren, um der automatischen Indexierung des klassischen Information Retrieval zu assistieren. Nach [Strzalkowski, 1995] scheint letzteres der bessere Ansatz zu sein, weil die erste Strategie sehr hohen Aufwand mit sich bringt.

Das folgende Zitat von [Lewis und Sparck Jones, 1996] verdeutlicht, welche Anforderungen dabei an die Beschreibung eines Textes gestellt werden: „DR [Document Retrieval] thus imposes conflicting demands on text descriptions, asking that they be normalizing, discriminating, and summarizing, as well as accurate“ (S.93). Zu ergänzen ist, daß Dokumentbeschreibungen auch so kurz wie möglich sein sollten, um lange Verarbeitungszeiten zu vermeiden. Die Information, die die morphologische Analyse mit GERTWOL bereitstellt, muß also gemäß der Anforderungen genutzt werden können.

GERTWOL ist ein System zur automatischen Wortformerkennung deutscher Wörter (vgl. [Haapalainen und Majorin, 1994]). Dabei kann es sowohl zur Analyse als auch zur Synthese deutscher Wörter eingesetzt werden. Entwickelt wurde GERTWOL bei der Firma LINGSOFT, INC. in Finnland, von der es auch vertrieben wird. Hier wird GERTWOL nur zur morphologischen Analyse herangezogen, und nicht zur Synthese genutzt.

Die theoretische Grundlage von GERTWOL bildet TWOL (TWO-Level model), eine sprachunabhängige morphologische Analysemethode, die von Prof. Dr. Kimmo Koskenniemi erarbeitet wurde. TWOL entstand im Rahmen eines Projektes mit dem Thema „Computeranalyse der finnischen Sprache“ an der Universität Helsinki unter Leitung von Prof. Dr. Fred Karlsson und wurde 1983 in [Koskenniemi, 1983b] abschließend dokumentiert. Die grundlegenden Prinzipien von TWOL werden in Abschnitt 4.1.1 näher beschrieben. GERTWOLs Eigenschaften und Leistungsgrenzen werden in 4.1.2 diskutiert. Auf welche Weise GERTWOL im Rahmen dieser Arbeit eingesetzt wird, erläutert Kapitel 4.1.3.

4.1.1 TWOL

Koskenniemi selbst bezeichnet TWOL in [Koskenniemi, 1983a] als „Two-level model for morphological analysis“. Bevor näher auf das Modell eingegangen wird, werden die beiden linguistischen Disziplinen, die Phonologie und die Morphologie, die der morphologischen Analyse durch TWOL zugrundeliegen, vorgestellt.

[Bünting, 1987] nennt eine allgemeinverständliche Umschreibung für die Sprachwissenschaft: Linguistik ist die „Lehre von den Lauten, ihrer Kombinierbarkeit zu Wörtern und deren Kombinierbarkeit zu Sätzen“. (Er merkt allerdings an, daß mit „solcher Umschreibung .. die Thematik linguistischer Bemühungen .. zu eng gefaßt ist“. Hier soll diese

Umschreibung aber ausreichen.)

Die morphologische Analyse stützt sich auf die Lehre von den Lauten und ihre Kombinierbarkeit zu Wörtern, d.h. sie nutzt die Theorien der *Phonologie* und der *Morphologie*.

Nach [Lyons, 1995] befaßt sich die *Morphologie* mit der Struktur der Wörter, die bestimmt wird von den Flexions- und Derivationsregeln einer Sprache. „In den Grammatiken der einzelnen Sprachen beschreibt der Abschnitt über Flexion die Deklinationen der Substantive, Adjektive und Pronomina und die Konjugationen der Verben [..]. Der Abschnitt über Derivation führt gewöhnlich verschiedene Verfahren an, wie aus bereits bestehenden Wörtern neue Wörter gebildet werden, z.B. Adjektive aus Nomina (‘triumphal’ aus ‘Triumph’), Nomina aus Verben (‘Hörer’ aus ‘hören’), Adjektive aus Verben (‘akzeptierbar’ aus ‘akzeptieren’) usw.“ ([Lyons, 1995], S.198f).

Die *Phonologie* dagegen ist nach [Lewandowski, 1975] „die Sprachgebildelehre, die sich [..] mit dem funktionierenden System der Laute“ befaßt. D.h. die Phonologie analysiert die Lautfolgen von Wörtern einer Sprache. Aus den Analyseergebnissen leitet sie Gesetzmäßigkeiten ab, nach denen Laute in einer Sprache zu Ausdruckselementen kombiniert werden. Diese Ausdruckselemente wiederum bestimmen die Struktur der Wörter.

Im Rahmen der generativen Transformationsgrammatik, die Chomsky erstmals 1957 in der Veröffentlichung von „Syntactic Structures“ vorstellte, werden Transformationsregeln formuliert, die einen phonologischen Prozeß beschreiben. Ein Beispiel für eine phonologische Transformationsregel im Deutschen ist:

$$i \rightarrow a/[- + Nasal + Konsonant]_{Vs2} + Prät.$$

(Das $-$ ist ein Platzhalter für den betreffenden Buchstaben.) Die Regel besagt, daß i durch a vor der Präteritumendung ersetzt wird, wenn es sich um Vs2-Stämme handelt, die auf Nasal (n) plus Konsonant (g) enden ($= sing$ ($< singen$) $\rightarrow sang$).

Koskenniemi bezieht sich in [Koskenniemi, 1983a] auf diese Regeln unter dem Begriff *generative Phonologie*. Erste Arbeiten konzentrieren sich darauf, die phonologischen Regeln in Computercode zu übersetzen. Diese Herangehensweise liegt besonders nahe, wenn man die Beschaffenheit der Regeln bedenkt, die [Bünting, 1987] beschreibt: „Wer unvorbereitet mit einer Transformationsgrammatik konfrontiert wird, der mag glauben, ein mathematisches Werk und nicht eine Grammatik in der Hand zu haben. Die Formeldarstellung der Regeln soll gewährleisten, daß bei der Beschreibung sprachlicher Mechanismen nicht an die Intuition und ein heuristisches Verstehen von Lesern appelliert wird, sondern daß durch die Beschreibung sprachliche Vorgänge explizit gemacht sind; je komplexer sprachliche Vorgänge sind, desto komplexer sind die Formeln.“

Der Formalismus der generativen Phonologie weist zwei Eigenschaften auf, die den Anlaß gaben, einen neuen Formalismus zu entwickeln. Zum einen basiert die generative Phonologie auf Transformationsregeln, die nur die beiden Operationen „*Substitution* (= Tilgen und Einsetzen von Elementen an gleicher Stelle) und *Permutation* (= Tilgen an einer und Einsetzen an anderer Stelle)“ ([Bußmann, 1983], S.551) erlauben. Dadurch wird eine Vorschrift bezüglich der Reihenfolge der Regelanwendungen erforderlich. Zum anderen ist die generative Phonologie nur in eine Richtung, zur Produktion von Wortformen, konzipiert.

Der Two-level-Formalismus basiert zwar zum Teil auf den gleichen Konzepten wie die generative Phonologie, versucht aber die oben genannten Eigenschaften zu vermeiden (vgl. [Koskenniemi, 1983b]).

Die Grundidee ist die Unterscheidung zweier Ebenen: der *lexikalischen Ebene* (lexical level) und der *Oberflächenebene* (surface level). Die *Oberflächenebene* bilden die zu untersuchenden Wörter, die wahlweise als Folge von Phonen (Lauten), als Folge von Phonemen (kleinste bedeutungsunterscheidende Einheiten) oder als Buchstabenfolge angegeben werden, je nach dem, was untersucht werden soll. Zur morphologischen Analyse erfolgt die Eingabe der Wörter als Buchstabenfolgen.

Die Repräsentation der Wörter auf der lexikalischen Ebene besteht aus dem Wortstamm und den erlaubten Endungen des Wortes. Diese Repräsentationen werden in einem Lexikon gesammelt, das die morphologischen Derivations- und Flexionsregeln enthält. Hinzu kommen noch einige phonologische Merkmale.

Regeln, die im wesentlichen die phonologischen Theorien formulieren, verbinden beide Ebenen. Im Gegensatz zu den Transformationsregeln der generativen Phonologie sind die Regeln so formuliert, daß sie nur vergleichen. Sie überprüfen, mit welchen Einträgen im Lexikon das zu untersuchende Wort auf der Oberflächenebene übereinstimmt. [Koskenniemi, 1983b] vergleicht diese Regeln mit einem mathematischen Gleichungssystem, wobei jede einzelne Regel eine Gleichung darstellt und die Wörter auf der Oberflächenebene und die Wörter auf der lexikalischen Ebene die Werte für die Variablen sind. Zur Erfüllung des Gleichungssystems müssen alle einzelnen Gleichungen mit den gegebenen Werten (und damit alle Regeln) erfüllt werden. Da die Regeln nicht selbst aktiv sind, d.h. Wörter bilden oder analysieren, sondern nur überprüfen, ist der Formalismus nicht nur auf eine Richtung beschränkt. Die Regeln werden in Endliche Automaten übersetzt, wobei teilweise mehrere Regeln in einem Automaten zusammengefaßt werden. Auf diese Weise wird verhindert, daß sich große Teile eines Automaten in anderen wiederholen.

Die oben beschriebenen Konzepte von Twol, die Trennung von Oberflächenebene und lexikalischer Ebene und die Idee der Gestaltung des Regelsystems als vergleichende Instanz, realisieren den sprachunabhängigen Teil des Modells. Um das Modell für unterschiedliche Sprachen anwendbar zu machen, müssen die zwei Hauptkomponenten des Systems, das *Lexikon* und das *Regelsystem* (Two-level rules), für die einzelnen Sprachen formuliert werden. Im weiteren werden die beiden Komponenten beispielhaft für das Finnische vorgestellt, denn die Beispiele sind aus [Koskenniemi, 1983b] und [Koskenniemi, 1983a] entnommen und beziehen sich auf das Finnische.

Das Lexikon Das Lexikon besteht aus einer Menge von Teillexika und einer Menge von Fortsetzungsklassen, wobei jedes Teillexikon und jede Fortsetzungsklasse eine Bezeichnung hat. Die Lexikoneinträge setzen sich aus drei Angaben zusammen:

1. Der *Wortstamm* wird je nach Anwendung als Folge von Phonen, als Folge von Phonemen oder als Buchstabenfolge repräsentiert.

2. Die *Fortsetzungsklasse* enthält die Namen der Teillexika, die alle möglichen Endungen angeben, mit denen der Stamm fortgesetzt werden kann.

3. Die *Wortinformation* beinhaltet alle Angaben, die nach der Identifikation eines Wortstamms oder einer Endung entschieden werden können. Zum Beispiel kann aus der Endung eines Substantivs Kasus und Numerus abgeleitet werden. Ob es sich hierbei um morphologische, syntaktische oder semantische Information handelt, entscheidet das Einsatzgebiet.

Ein Beispiel für einen Lexikoneintrag eines Wortstamms lautet:

`katTo /S „Roof S“`

`katTo` gibt die phonologische Repräsentation des Stammes an. `/S` bezeichnet die Fortsetzungsklasse und „`Roof S`“ liefert die ableitbare Wortinformation (hier die englische Übersetzung und das Kürzel für die Wortart Substantiv). Der Eintrag für die Fortsetzungsklasse `/S` lautet wiederum

`katTo /S = S0 S1 S2 S3,`

wobei S0, S1, S2 und S3 Teillexika sind, die mögliche Endungen für den Stamm `katTo` enthalten.

Wenn zur Analyse das Wort `katto` auf der Oberflächenebene angegeben wird, wird im Lexikon nach Einträgen gesucht, so daß die Wortstämme auf beiden Ebenen übereinstimmen. Im Beispiel findet sich also `katTo /S` „`Roof S`“. Damit steht fest, daß es sich um ein Substantiv (`S`) mit der englischen Übersetzung „`Roof`“ handelt. Danach werden die möglichen Endungen dieses Wortstamms in den angegebenen Fortsetzungsklassen untersucht. Wird eine übereinstimmende Endung gefunden, wird dem entsprechenden Eintrag die fehlende Information entnommen. Die Übereinstimmung zwischen der Oberflächenebene und der lexikalischen Ebene muß keine Gleichheit bedeuten. Es sind auch Abweichungen erlaubt. Welche Paarbildungen zwischen Oberflächenebene und lexikalischer Ebene zulässig sind, beschreibt das Regelsystem.

Das Regelsystem Der Aufbau der Regeln soll an einem Beispiel erklärt werden. Die Gleichung

$$\genfrac{}{}{0pt}{}{\mathtt{i}}{\mathtt{j}} <\equiv> \mathtt{V} + - \mathtt{V}$$

formuliert die finnische Sprachregel, daß das `i`, das den Plural einiger Substantive bezeichnet, durch `j` ersetzt wird, wenn es zwischen zwei Vokalen `V` steht. Die übereinander angeordnete Schreibweise von `i` und `j` deutet die zwei Ebenen an. Die obere Zeile ist die lexikalische Ebene, die untere ist die Oberflächenebene. Das − ist wiederum Platzhalter für den Buchstaben, auf den sich die Regel bezieht, und das + ist ein Konjunktor, der den Stamm mit der Endung verbindet.

Wie oben bereits angeführt, können die Regeln entweder zur Produktion oder zur Analyse eingesetzt werden. Wird der Stamm eines Wortes und die gewünschte Wortform angegeben, kann die Darstellung des Wortes auf der lexikalischen Ebene ermittelt und daraus das gesuchte Wort auf der Oberflächenebene bestimmt werden. In der anderen Richtung wird das zu analysierende Wort in der Darstellung der Oberflächenebene darauf untersucht, mit welchen Lexikonregeln es abgeglichen werden kann. Um die Vorgehensweise des Regelsystems deutlicher zu machen, nimmt man beide Darstellungen als gegeben an:

```
Lexical level:   t a l o + i A
Surface level:   t a l o ∅ j a
```

Da die beiden Ebenen buchstabenweise miteinander verglichen werden, werden Nullzeichen (∅) eingefügt, so daß korrespondierende Buchstaben übereinander stehen. In diesem

Beispiel trifft die oben genannte Regel zu, daß das `i` zwischen zwei Vokalen (hier `o` und `a`) zu `j` wird.

Die Regeln sind als Endliche Automaten implementiert. Die Eingabe sind Symbolpaare: je ein Symbol der lexikalischen Ebene und ein Symbol der Oberflächenebene werden miteinander verglichen, wobei die Kombination ausschlaggebend für den nächsten Zustand ist. Abbildung 2 zeigt den Automaten für die oben angeführte Plural-`i`-Regel (das Gleichheitszeichen steht hier für einen beliebigen Konsonanten, während das V Vokale symbolisiert).

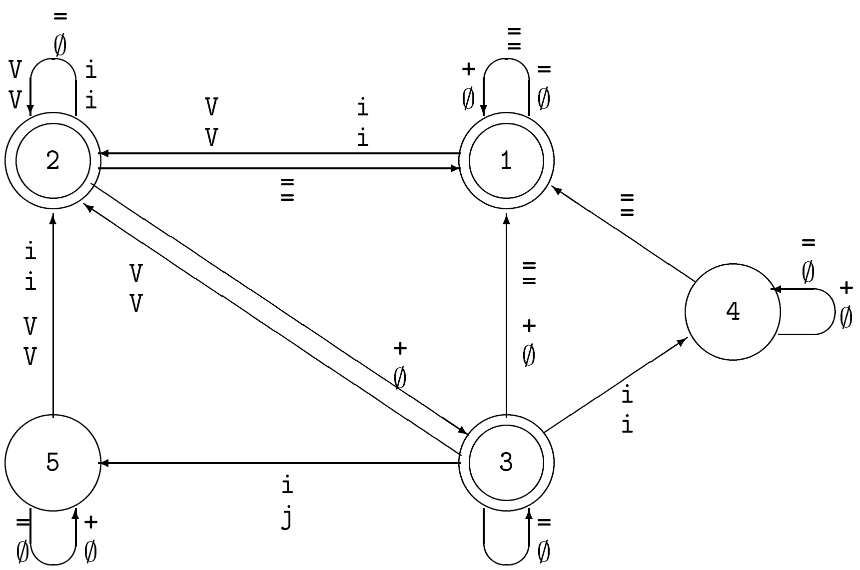

Abbildung 2: Endlicher Automat für die Plural-`i`-Regel

Zustand `1` ist der Initialzustand. Terminale Zustände sind durch zwei konzentrische Kreise markiert, d.h. in diesem Fall, wenn der Automat in Zustand `4` oder `5` endet, wird die Regel verworfen.

Wendet man den Automaten auf die Beispieleingabe an, werden folgende Zustände durchlaufen:

```
Lexical:    t a l o + i A
Surface:    t a l o Ø j a
State:     1 1 2 1 2 3 5 2
```

Das Regelsystem wirkt wie ein Filter. Läßt sich das Wort der Oberflächenebene nicht durch Anwendung einer Regel mit einem Lexikoneintrag in Übereinstimmung bringen, wird dieser Eintrag verworfen. Andererseits werden alle Lexikoneinträge, die zutreffen, ausgegeben. So liefert das System auch bei homographischen Wortformen alle korrekten Interpretationen.

4.1.2 Eigenschaften von Gertwol

Gertwol verfügt über ein großes Lexikon, dem das komplette Sprachmaterial des Deutsch-Englischen Wörterbuchs von Collins (The Collins German Dictionary, Neubearbeitung 1991) zugrunde liegt. Nach Tests an unterschiedlichen Corpora wurde das Lexikon noch ergänzt, so daß sich die Gesamtzahlen nach [Haapalainen und Majorin, 1994] jetzt auf folgende Werte belaufen:

- 11.000 Adjektive
- 2.000 Adverbien
- 400 Interjektionen
- 50.000 Substantive
- 6.500 Verben
- 12.000 Eigennamen
- 1.700 Abkürzungen

Hinzu kommen Konjunktionen, Pronomina, Artikel und Präpositionen. Durch eine umfassende Derivationsmorphologie und einen vollständigen Mechanismus zur Kompositabildung erreichte Gertwol bei Tests an unterschiedlichen Corpora (u.a. an Zeitungstexten) eine Datenabdeckung von 99% bei orthographisch einwandfreien Texten und über 98% bei unbeschränkten Texten (vgl. auch [Haapalainen und Majorin, 1994]). Nicht analysierte Wörter werden in der Ausgabe kenntlich gemacht, so daß sie bei der Weiterverarbeitung berücksichtigt werden können.

Bei Eingabe eines Wortes oder eines Textes analysiert Gertwol jedes Wort einzeln ohne Berücksichtigung des Kontextes, in dem das Wort steht. Es gibt zu jeder eingegebenen Wortform die Grundform, die Wortart und morphologische Daten zum Numerus, Genus, Kasus, Modus und Tempus sowie zur Person und Komparation an.

Zu dem Beispielsatz „Bürgermeister duldet keine Duldung.“ (Schlagzeile der TAZ vom 23.07.1996) liefert GERTWOL folgende Ausgabe:

```
"<*buergermeister>"
        "*buerger#meister"  S MASK SG NOM
        "*buerger#meister"  S MASK SG AKK
        "*buerger#meister"  S MASK SG DAT
        "*buerger#meister"  S MASK PL NOM
        "*buerger#meister"  S MASK PL AKK
        "*buerger#meister"  S MASK PL GEN

"<duldet>"
        "duld~en"  V IND PRAES PL2
        "duld~en"  V IMP PRAES PL2
        "duld~en"  V KONJ PRAES PL2
        "duld~en"  V IND PRAES SG3

"<keine>"
        "kein"  PRON INDEF NEG SG NOM FEM
        "kein"  PRON INDEF NEG SG AKK FEM
        "kein"  PRON INDEF NEG PL NOM
        "kein"  PRON INDEF NEG PL AKK
        "kein"  DET INDEF NEG SG NOM FEM
        "kein"  DET INDEF NEG SG AKK FEM
        "kein"  DET INDEF NEG PL NOM
        "kein"  DET INDEF NEG PL AKK

"<*duldung>"
        "*duld~ung"  S FEM SG NOM
        "*duld~ung"  S FEM SG AKK
        "*duld~ung"  S FEM SG DAT
        "*duld~ung"  S FEM SG GEN

"<--punkt>"
        ""  PUNKT
```

<*buergermeister> wird eindeutig als Substantiv (S) im Maskulinum (MASK) erkannt, wobei der Numerus, also Singular (SG) oder Plural (PL), und der Kasus nicht bestimmt werden können. Das Verb (V) <duldet> mit der Grundform „dulden“ steht im Präsens (PRAES), aber über den Modus des Verbs, also Indikativ (IND), Konjunktiv (KONJ) oder Imperativ (IMP), oder über die Person (2.Person Plural/PL2 oder 3.Person Singular/SG3) läßt sich keine Aussage machen. <keine> ist ein verneintes (NEG) Indefinitpronomen (PRON INDEF) oder ein Determinativpronomen (DET) mit der Grundform „kein“, wobei ebenfalls der Kasus und der Numerus nicht eindeutig bestimmt werden können. Bei dem Substantiv <*duldung> im Femininum Singular (SG FEM) stimmen eingegebene Form und Grundform überein. Der Kasus ist allerdings nicht eindeutig.

Die Sonderzeichen `*`, ~ und `#` haben folgende Bedeutung: `*` kennzeichnet Großschreibung, ~ trennt Suffixe ab und `#` stellt starke Kompositagrenzen dar. Darüber hinaus markiert GERTWOL auch schwache Kompositagrenzen (`|`) und Fugenelemente (\).

Problematisch an der Ausgabe von GERTWOL ist die Vielzahl ermittelter Interpretationen für ein eingegebenes Wort.

Zum einen entsteht diese Vielfalt dadurch, daß sich viele Wortformen im Deutschen nicht voneinander unterscheiden. Erst im Zusammenhang mit den Wörtern des ganzen Satzes kann ein Wort des Satzes genauer analysiert werden. Z.B. kann der Kasus eines Substantivs unter Berücksichtigung des zugehörigen Artikels enger eingegrenzt werden. Die Benutzung einer Grammatik, die diese syntaktischen Merkmale enthält, ist bei GERTWOL allerdings nicht vorgesehen.

Außerdem entstehen besonders bei langen, zusammengesetzten Wörtern verschiedene Interpretationen durch unterschiedliche Zerlegungsmöglichkeiten der Wörter. Als Beispiel sei hier die Zerlegung des Wortes „Hausaufgabe“ genannt:

```
"<*hausaufgabe>"
"*haus#auf|gab~e"  S FEM SG NOM
"*haus#auf|gab~e"  S FEM SG AKK
"*haus#auf|gab~e"  S FEM SG DAT
"*haus#auf|gab~e"  S FEM SG GEN
"*hau#sauf#gab~e"  S FEM SG NOM
"*hau#sauf#gab~e"  S FEM SG AKK
"*hau#sauf#gab~e"  S FEM SG DAT
"*hau#sauf#gab~e"  S FEM SG GEN
```

Da GERTWOL auch keine semantischen Merkmale nutzt, können unsinnige Zerlegungen wie „`*hau#sauf#gab~e`“ auch nicht ausgeschlossen werden.

4.1.3 Einsatzbereich

Mit GERTWOL beginnt die Bearbeitung der Zeitungsartikel. An dieser Stelle wird im klassischen Information Retrieval damit begonnen, den Text der Eingabe, der zunächst nur eine Aneinanderreihung von Zeichen ist, in Wörter zu zerlegen. Diese Vorverarbeitung leistet hier ebenfalls GERTWOL. Die ungekürzten Texte werden Wort für Wort morphologisch analysiert. Für das im Anschluß an die Analyse folgende automatische Indexieren werden die von GERTWOL ermittelten Grundformen genutzt. Außerdem wird die Information berücksichtigt, um welche Wortarten es sich bei den Wörtern der Artikel handelt.

Aufgrund der oben angeführten Probleme, daß bei der Analyse eines Wortes mehrdeutige Interpretationen möglich sind, wurde darauf verzichtet, morphologische Daten, die nicht eindeutig von GERTWOL bestimmt werden können, zu verwenden. Um diese Daten zu nutzen, müssen mehrdeutige Interpretationen disambiguiert werden. Dazu muß der morphologischen Analyse eine syntaktische Analyse folgen. Die Firma LINGSOFT, INC. arbeitet an einem Werkzeug zur Disambiguierung, das aber bis jetzt noch nicht fertiggestellt wurde.

4.2 Automatische Indexierung

Ziel der Indexierung im Information Retrieval ist es, eine charakterisierende Darstellung für ein Dokument zu finden, die es erlaubt bei einer Nachfrage zu testen, ob das Dokument in Bezug auf die Nachfrage interessant ist oder nicht. Außerdem soll die durch Indexieren ermittelte Darstellung des Dokumentes einen Vergleich der Dokumente untereinander ermöglichen. Dazu werden zu jedem Dokument beschreibende Schlüsselwörter ermittelt. [Jüttner und Güntzer, 1988] unterscheiden *manuelle Indexierung*, bei der ein Schlüsselwortverzeichnis von Fachkräften verwaltet wird, und *automatische Indexierung*, die ohne menschliche Unterstützung vorgenommen wird. Die manuelle Indexierung ist für die meisten Anwendungen nicht praktikabel, weil sie zum einen zu zeitaufwendig ist, und weil zum anderen nach [Ingwersen, 1992] bei der Indexierung durch verschiedene Fachkräfte Inkonsistenzen auftreten.

Die verschiedenen Ansätze für Verfahren zur automatischen Indexierung werden in Abschnitt 4.2.1 vorgestellt. Außerdem begründet dieser Abschnitt die Wahl der Verfahren für das System AKAT. Abschnitt 4.2.2 beschreibt die Systemkomponente zur automatischen Indexierung.

4.2.1 Verfahren

Christopher Fox definiert automatisches Indexieren in [Fox, 1992] mit folgenden Worten:

> „*Automatic indexing* is the process of algorithmically examining information items to generate lists of index terms.“ (S.102)

Mit „lists of index items“ bezeichnet Fox die gesuchte charakterisierende Darstellung der zu indexierenden Dokumente. Dabei muß die Dokumentdarstellung zwei Anforderungen erfüllen: Erstens muß sie so strukturiert sein, daß sie Vergleiche zwischen Anfrage und Dokument bzw. zwischen zwei Dokumenten ermöglicht, und zweitens muß sie den Inhalt des Dokumentes repräsentieren. Ulrich Reimer spricht in [Reimer, 1992] statt von „automatischer Indexierung“ auch von „automatischer Inhaltserschließung“ und unterscheidet vier verschiedene Ansätze:

- zeichenkettenorientierter,
- statistischer,
- linguistischer und
- wissensbasierter Ansatz.

Den Verfahren aller vier Ansätze geht eine Vorverarbeitung voraus, die die Texte in einzelne Wörter zerlegt und sogenannte Stoppwörter eliminiert. Stoppwörtern wird keine Relevanz für die Charakterisierung eines Textes beigemessen, weil sie zu allgemein sind und dadurch keinen Rückschluß auf den Inhalt des Textes zulassen. Diese Wörter werden zu einer Stoppwortliste zusammengefaßt, mit der die Dokumente gefiltert werden. Beispiele für Stoppwörter im Deutschen sind bestimmte und unbestimmte Artikel, Konjunktionen, Pronomen, Modalverben u.ä. Die automatische Indexierung wird auf die so verkürzten Texte angewandt.

Bei einer solchen, stark wortbezogenen Analyse der Texte zur Inhaltserschließung ergeben sich verschiedene Probleme:

- Der Wortschatz einer Sprache umfaßt viele Wörter mit gleicher Schreibweise aber unterschiedlicher Bedeutung, z.B. „Bank“ und „Tenor“.
- In einer Sprache lassen sich verschiedene Wörter finden, um denselben Sachverhalt auszudrücken, z.B. „Auto“, „Wagen“ und „PKW“.
- Besonders flexionsreiche Sprachen, wie Deutsch, erschweren das automatische Indexieren, weil viele Derivations- und Flexionsformen und Komposita die Bandbreite der unterschiedlichen Formulierungen noch erweitern. Z.B. haben deklinierte Formen wie „Haus“, „Hauses“ und „Häuser“ für die Inhaltserschließung dieselbe Bedeutung.
- Die Wörter, die in einem Text vorkommen, müssen für den Textinhalt nicht alle gleich aussagekräftig sein. Die Wortwahl eines im Text angeführten Beispiels oder eines aufgezeigten Gegensatzes ist zur Inhaltserschließung meistens weniger relevant als der übrige Text.

Die vier Ansätze gehen auf diese Probleme unterschiedlich ein: für die Verfahren, die auf dem *zeichenkettenorientierten Ansatz* beruhen, ist ein Text nur eine Folge von Wörtern, die wiederum eine Folge von Zeichen sind. Entscheidend ist nur, ob ein Wort in einem Text zu finden ist, oder nicht.

Um das Problem der Flexion, Derivation und Kompositabildung zu bewältigen, werden Zeichenkettenoperatoren eingesetzt, mit denen eine generalisierte Wortform angegeben werden kann, die alle Formen eines Wortes abdecken soll (z.B. steht „h$$s#“ für „Haus“, „Hauses“ und „Häuser“, aber auch für „Hanse“ oder „hassen“). Aus dem Beispiel ist ersichtlich, daß der Einsatz von Zeichenkettenoperatoren das Problem nicht zufriedenstellend löst.

Weitere Nachteile dieses Ansatzes ergeben sich aus der Gleichbehandlung aller Wörter im Text: Es wird nicht unterschieden, daß ein Wort häufiger vorkommt als ein anderes, ob also ein Wort für ein Dokument relevanter ist als ein anderes. Außerdem können doppeldeutige Wörter nicht differenziert werden und in der Wortwahl unterschiedliche Formulierungen nicht auf einen Begriff zurückgeführt werden.

Die Verfahren des *statistischen Ansatzes* setzen an einer Schwachstelle des zeichenkettenorientierten Ansatzes an: Sie kontrollieren nicht nur das Auftreten eines Wortes, sondern sie streben auch eine Differenzierung der Wörter nach deren Relevanz in Hinblick auf den Textinhalt an. Dazu weisen sie jedem Wort ein Gewicht zu, das die Relevanz des Wortes für den Text widerspiegelt. Folgende Beobachtungen liegen der Gewichtung zugrunde:

- Ein Wort ist umso aussagekräftiger zur Inhaltserschließung eines Dokumentes, je häufiger das Wort in diesem Dokument auftritt.
- Ein Wort ist weniger aussagekräftig für die Inhaltserschließung, wenn es gleichverteilt in vielen Dokumenten vorkommt, während ein Wort, das nur in wenigen Texten vorkommt, signifikant für diese Texte ist.

Bei der Berechnung der Gewichte gehen diese Beobachtungen in die Bestimmung der Faktoren tf und idf ein, die in Abschnitt 4.2.2 näher erläutert werden.

Im Gegensatz zum zeichenkettenorientierten Ansatz, der jeden Text für sich betrachtet, setzt der statistische Ansatz die zu bearbeitenden Dokumente also zueinander in Beziehung. Da die statistischen Verfahren genauer zwischen vorkommenden Wörtern differenzieren, stellen sie eine wesentliche Verbesserung gegenüber den zeichenkettenorientierten Verfahren dar. Trotzdem bleibt auch hier das Problem, daß bei stark flektierten Sprachen verschiedene Flexions- und Derivationsformen eines Wortes nicht zu einem Wort zusammengeführt werden und Mehrdeutigkeiten nicht aufgelöst werden können.

Unter dem *linguistischen Ansatz* faßt [Reimer, 1992] die Verfahren zusammen, die vor der eigentlichen Indexierung eine morphologische und/oder eine syntaktische Analyse einsetzen, die auf linguistischem Wissen beruhen. Dagegen stellt er Verfahren des wissensbasierten Ansatzes, die diskursspezifisches Wissen (nach [Reimer, 1992] = semantisches Wissen + Weltwissen) voraussetzen.

[Fuhr, 1995] unterscheidet bei den linguistischen Verfahren graphematische und lexikalische. Wie die zeichenkettenorientierten Verfahren betrachten graphematische Verfahren die Wörter als Buchstabenfolgen. Aber anstatt der Zeichenkettenoperatoren werden Regeln verwendet, nach denen sich aus den Buchstabenfolgen die Grundformen der Wörter ableiten lassen. Der Einsatz graphematischer Verfahren empfiehlt sich bei schwach flektierten Sprachen wie Englisch, die mit wenigen Regeln auskommen. Lexikalischen Verfahren liegt ein Lexikon zugrunde, das Derivations-, Flexionsformen und Komposita der Wörter einer Sprache enthält. Die morphologische Analyse erlaubt es, durch Reduktion flektierter Wörter auf eine Grundform, unterschiedlichen, grammatikalischen Formulierungen das gleiche Wort zuzuweisen.

Eine syntaktische Analyse in Form eines partiellen (Konzentration auf einzelne Satzteile wie Nominalphrasen) oder eines vollständigen Parsings (Untersuchung ganzer Sätze) ermöglicht es, die Wörter stärker zueinander in Beziehung zu setzen. Z.B. kann der Ausdruck „die Tür des Hauses“ dem Begriff „Haustür“ gleichgesetzt werden.

Als nachteilig erweist sich auch beim linguistischen Ansatz, daß mehrdeutige Wörter und variierende Wortwahl nicht unterschieden werden können. Außerdem reicht selbst eine vollständige, syntaktische Satzanalyse nicht aus, um die für eine Indexierung wesentlichen Begriffe zu identifizieren (vgl. auch [Reimer, 1992]).

„Der methodische Kern des *wissensbasierten Ansatzes* zur Volltextverarbeitung besteht aus den Wissensrepräsentations- und -transformationskonzepten der Künstlichen Intelligenz“ ([Hahn, 1986], S.204). Ziel ist es, eine Repräsentation des Dokumentes zu finden, die unabhängig von der sprachlichen Formulierung ist. Dazu werden Wissensrepräsentationsformate wie semantische Netze oder Frames eingesetzt. Verfahren dieses Ansatzes sind meistens mächtiger als die oben beschriebenen Verfahren zur Inhaltserschließung von Dokumenten, weil die verwendeten Repräsentationssprachen den Aufbau und die Ergänzung einer Wissensbasis erlauben, die es ermöglicht, das Textverstehenssystem zu einem Frage-Antwort-System oder Textzusammenfassungssystem zu erweitern.

Wie oben bereits erwähnt, setzt die Anwendung wissensbasierter Verfahren allerdings Information über einen Diskursbereich voraus. Ein einfaches Verfahren stellt der Einsatz eines Thesaurus dar, der verschiedene Synonyme auf eine Bezeichnung abbildet. Verfeinert wird diese Idee durch eine Begriffshierarchie, die nicht nur äquivalente Begriffe zueinander in Beziehung setzt, sondern auch Ober- und Unterbegriffe einführt. Damit können Variationen in der Wortwahl eines Textes aufgefangen werden. Außerdem kann dem Lexikon eine semantische Komponente hinzugefügt werden, die Mehrdeutigkeiten auflöst. Der

Nachteil dieser Verfahren ergibt sich aus der Spezialisierung auf einen Diskursbereich.

Die Differenzierung zwischen linguistischen Verfahren und wissensbasierten Verfahren, die [Reimer, 1992] vornimmt, entspricht nicht dem Ansatz in den Abschnitten 2.1 und 2.2, in denen alle Methoden, die Wissen über Sprache verwenden, unter dem Begriff NLP-Techniken zusammengefaßt werden. Nach dieser Zusammenfassung gehören die Verfahren beider Ansätze zu Verfahren zur automatischen Indexierung, die NLP-Techniken einsetzen.

Ein weiterer Ansatz, den Inhalt eines Dokumentes zu erschließen, ist die Berücksichtigung der Struktur des Dokumentes. Wissenschaftliche Aufsätze z.B. sind meist ähnlich gegliedert. Sie beginnen mit einer Zusammenfassung, die einen kurzen Überblick über den Aufsatz geben soll. Es liegt nahe, daß die Zusammenfassung sehr viele für das Aufsatzthema wichtige Schlüsselwörter enthält. Durch stärkere Gewichtung der Wörter der Zusammenfassung kann diese Beobachtung berücksichtigt werden. [Ingwersen, 1992] führt eine Methode an, die auf einer ähnlichen Idee beruht und sich ebenfalls auf wissenschaftliche Arbeiten bezieht: Um den Inhalt eines Dokumentes automatisch zu erschließen, wird nur der Titel auf Schlüsselwörter untersucht, mit der Begründung, daß die Titel wissenschaftlicher Arbeiten meist aussagekräftig formuliert werden.

Die bis hierhin aufgeführten Ansätze mit ihren Vor- und Nachteilen und die Besonderheiten des Anwendungsgebietes leiten die Wahl der Verfahren, die hier eingesetzt werden:

Der zeichenkettenorientierte Ansatz geht nur zu einem kleinen Teil auf die bei der automatischen Indexierung entstehenden Probleme ein und bietet auch zur Behandlung von Flexion, Derivation und Kompositabildung keine zufriedenstellende Lösung an. Damit wurde dieser Ansatz hier nicht weiter verfolgt.

Die Nachteile des statistischen Ansatzes werden durch Kombination mit Verfahren des linguistischen Ansatzes kompensiert. Da die hier vorgestellte Anwendung auf Zeitungsartikeln in deutscher Sprache arbeitet, können graphematische Verfahren, die sich bei flexionsarmen Sprachen wie Englisch gut bewährt haben, nicht eingesetzt werden. Deshalb wird die morphologische Analyse hier auf der Grundlage eines umfassenden Lexikons der deutschen Sprache vorgenommen. Ziel der morphologischen Analyse ist die Reduktion der flektierten Wörter auf eine Grundform, um eine präzisere Gewichtung der im Dokument vorkommenden Wörter zu erzielen, und die Bestimmung der Wortarten.

Unter der Annahme, daß gerade im Bereich des Zeitungswesens einige Wortarten einen höheren Beitrag zur Inhaltserschließung eines Dokumentes leisten (vgl. 3.1), wird das Ergebnis der Wortartbestimmung genutzt, um aus einem Dokument weniger aussagekräftige Wortarten herauszufiltern. Damit wird das Problem umgangen, das durch die Derivationsmöglichkeiten der deutschen Sprache hervorgerufen wird, denn abgeleitete Wortarten müssen nicht mehr auf die Ausgangswortarten zurückgeführt werden, weil sie herausgefiltert werden. Die gezielte Entfernung von Wortarten wie Artikel, Pronomen u.ä. ersetzt außerdem die Stoppworteliminierung.

Der Verzicht auf eine syntaktische Analyse und auf Verfahren des wissensbasierten Ansatzes, die eine Einschränkung des Diskursbereiches nach sich ziehen, wurde bereits in Abschnitt 3.2 begründet.

Ob die Gewichtung von Überschriften im Bereich Zeitungswesen ähnlich sinnvoll ist wie bei wissenschaftlichen Aufsätzen, muß sich in Experimenten zeigen. Nach den Ausführungen zur besonderen Syntax von Artikelüberschriften in Abschnitt 3.1 sind zwei Methoden zur Verkürzung von Überschriften üblich, Ersparungen und Auslassungen. Erstere erlauben noch den Rückschluß auf das Artikelthema und sprechen für die Gewichtung der Überschriften, während Auslassungen das Thema verschleiern und sich nicht unbedingt zur Inhaltserschließung eignen. Der folgende Abschnitt befaßt sich näher mit den Verfahren, die in der Systemkomponente zur automatischen Indexierung zum Einsatz kommen.

4.2.2 Die Komponente zur automatischen Indexierung

Die automatische Indexierung folgt der morphologischen Analyse mit GERTWOL. Zu Beginn werden die von GERTWOL erzeugten Ergebnisse gefiltert, so daß die resultierenden Dokumente nur noch aus den Grundformen bestehen, auf die GERTWOL die im Originaltext vorkommenden Wortformen reduziert hat. Gleichzeitig wird an dieser Stelle die Filterung nach Wortarten vorgenommen. In Abschnitt 4.2.1 wurde darauf hingewiesen, daß es üblich ist, aus Dokumenten Stoppwörter zu eliminieren. Es ist aufwendig, eine Stoppwortliste manuell zu erstellen. Außerdem kann die Vollständigkeit einer solchen Liste nicht garantiert werden. Eine Alternative bietet hier also die Filterung nach Wortarten. Durch die Einschränkung auf sehr wenige Wortarten geht die Filterung über die übliche Stoppworteliminierung noch hinaus. Kapitel 5 erläutert, ob die Verkürzung der Texte durch Wortartauswahl die Clusteranalyse verbessert. Außerdem beantwortet es die Frage, wie weit Texte verkürzt werden können, ohne daß der Informationsverlust zu hoch ist, um ein sinnvolles Clustering durchzuführen.

Dieser Abschnitt wird damit fortgesetzt, die abstrakten Überlegungen des vorhergehenden Abschnitts umzusetzen. Diese Überlegungen zur automatischen Generierung der charakteristischen Darstellung eines Dokumentes bezogen sich auf den Inhalt einer solchen Darstellung. Gleichzeitig muß diese Darstellung aber auch operational anwendbar sein. Dazu wird ein Dokument d als Vektor repräsentiert mit

$$\vec{d} = (t_a, t_b, \ldots, t_p),$$

wobei $t_a, t_b, \ldots, t_p$ die im Dokument d vorkommenden Wörter bezeichnen, die im weiteren in Anlehnung an die Literatur (siehe z.B. [van Rijsbergen, 1979]) *Terme* genannt werden. Die nach der Analyse durch GERTWOL gefilterten Artikel ergeben jeweils einen solchen *Dokumentvektor*.

Während der automatischen Indexierung wird den Termen t_i eines Dokumentes d_k ein Gewicht w_{i,d_k} zugewiesen. Die Berechnung des Gewichtes w_{i,d_k} leitet sich aus den folgenden, oben bereits angeführten Beobachtungen ab:

Terme, die häufig in einem Dokument erwähnt werden, sind signifikant für das Dokument. Damit sollte die *Termhäufigkeit* $tf_{i,k}$ eines Terms t_i in einem Dokument d_k zur Gewichtung berücksichtigt werden.

Zusätzlich muß zum Ausdruck kommen, daß ein Term, der in vielen Dokumenten vorkommt, an Signifikanz für das betrachtete Dokument d_k verliert. Die *Dokumenthäufigkeit* df_i des Terms t_i geht als *inverse Dokumenthäufigkeit* idf_i in die Berechnung ein:

$$idf_i := \log \frac{N}{n_i}$$

mit $N =$ Anzahl der zu indexierenden Dokumente, $n_i =$ Anzahl der Dokumente, in denen t_i vorkommt. Ein Term charakterisiert den Inhalt eines Dokumentes also dann am besten, wenn er eine hohe Termhäufigkeit und eine niedrige Dokumenthäufigkeit (und damit eine hohe inverse Dokumenthäufigkeit) aufweist. Durch Multiplikation dieser beiden Faktoren ergibt sich das Gewicht w_{i,d_k} dann als:

$$w_{i,d_k} := tf_{i,k} \cdot idf_i, \qquad k = 1 \ldots N, i = 1 \ldots t$$

Nach der Indexierung werden die Dokumente d_k also repräsentiert durch den Vektor:

$$\vec{d_k} = ((t_1, w_{1,d_k}); (t_2, w_{2,d_k}); \ldots; (t_t, w_{t,d_k}))$$

Da die Vektoren umfangreicherer Dokumente in der Regel länger sind als die Vektoren kurzer Dokumente, müssen die Dokumentvektoren normiert werden. Aus der Division der Vektoren durch ihre Beträge

$$|\vec{d_k}| = \sqrt{w_{1,d_k}^2 + w_{2,d_k}^2 + \ldots + w_{t,d_k}^2}$$

resultieren die normierten Dokumentvektoren.

Neben der beschriebenen Vorgehensweise zur Bestimmung der Gewichte finden sich in der Literatur noch weitere Ansätze (z.B. [Fuhr, 1995]), die sich in der Berechnung der Termhäufigkeit, der Dokumenthäufigkeit und bei der Normierung der Dokumentvektoren unterscheiden können. Welche Rechenvorschriften sich empfehlen, ist abhängig von den Dokumenten, was in [Salton und Buckley, 1988] nachgewiesen wird.

Gerard Salton und Chris Buckley haben in einer Studie jeweils drei verschiedene Ansätze zur Berechnung der Häufigkeiten miteinander kombiniert, wobei teilweise auch auf die Normierung verzichtet wurde, und die verschiedenen Kombinationen getestet. Aufgrund der Ergebnisse der experimentellen Vergleiche kommen sie u.a. zu folgenden Schlußfolgerungen:

Im Gegensatz zu Dokumenten, die sich durch technisches Vokabular und viele bedeutungsvolle Terme auszeichnen, empfehlen sie bei Dokumenten, die ein stark variierendes Vokabular aufweisen, das oben angeführte Maß für die Termhäufigkeit.

Außerdem sollte bei langen Dokumentvektoren ebenso wie bei extrem variierender Vektorlänge eine Normierung vorgenommen werden.

Da es sich bei den hier zur automatischen Indexierung vorliegenden Dokumenten um Zeitungsartikel handelt, treffen die Schlußfolgerungen zu: Zeitungsartikel zeichnen sich im allgemeinen nicht durch spezielles, technisches Vokabular aus, sondern variieren in der Wortwahl besonders aufgrund der weiten Themenbandbreite. Hinzu kommt, daß in Zeitungen nur in Ausnahmen spezielles, technisches Vokabular verwendet wird, wenn bei den Lesern ausreichendes Fachwissen und Verständnis vorausgesetzt werden kann.

Typisch für Zeitungen ist auch die Variation der Artikellängen. Häufig wechseln sich Kurzmeldungen, aktuelle Berichte oder lange Reportagen ab, oft in Abhängigkeit von den Ressorts, denen die Artikel zugeordnet werden.

4.3 Clusteranalyse

Ziel der Clusteranalyse ist es, Gruppierungen ähnlicher Objekte in einer ungeordneten Menge ausfindig zu machen. Diese Gruppierungen werden als „Cluster“ oder „Gruppen“ bezeichnet. [Panyr, 1986] verwendet auch den Begriff „Klasse“ synonym zu „Cluster“ und „Gruppe“, so wie er auch nicht zwischen den Begriffen „Klassifikation“, „automatische Klassifikation“, „Clusteranalyse“ und „Clustering“ unterscheidet.

Diese undifferenzierte Begriffsbildung hält [Willet, 1988] für irreführend und begründet:

> „Classification normally refers to the *assignment* of objects to predefined classes whereas cluster analysis requires the *identification* of these classes; thus clustering must precede classification in the analysis of a dataset.“ (S.577)

[Morik, 1995a] unterscheidet sogar drei Phänomenbereiche: die Kategorisierung (oder Aggregation), die Charakterisierung (oder Definition) und die Klassifikation. Während die Kategorisierung Objekte in Klassen gruppiert, d.h. zu Kategorien zusammenfaßt, setzt sich die Charakterisierung mit der Beschreibung der Kategorien auseinander. Die Klassifikation nutzt die Beschreibungen, um neue Objekte zu Kategorien zuzuordnen.

Die Clusteranalyse ist ein Verfahren zur Kategorisierung. Im Gegensatz zur Klassifikation kann sie aber nicht auf gegebene Kategorien und deren Beschreibungen zurückgreifen, sondern muß automatisch einander ähnliche Objekte zu einem Cluster zusammenführen und einander unähnliche Objekte trennen. Dazu ist ein Ähnlichkeitsmaß notwendig, mit dem zwei einander ähnliche Objekte von zwei einander unähnlichen Objekten unterschieden werden können.

Die Idee der Clusteranalyse wurde zunächst in der Biologie intensiv verfolgt. Inzwischen wird sie aber in vielen Bereichen, darunter die Medizin, die Archäologie, die Astronomie und die Geologie, um nur einige, wenige zu nennen, eingesetzt. Durch diese breit gestreuten Einsatzbereiche existieren sehr viele unterschiedliche Verfahren zur Clusteranalyse. Da diese Arbeit in den Bereich des Information Retrieval fällt, beschränken sich die folgenden Ausführungen auf Verfahren zur Unterstützung des Information Retrievals.

Im IR wird die Clusteranalyse eingesetzt, um die Effizienz und die Effektivität des Retrievals zu erhöhen, oder um Strukturen, die der Literatur eines Themengebietes zugrundeliegen, zu entdecken (vgl. [Rasmussen, 1992]). Dazu wurden bisher drei verschiedene Ansätze verfolgt:

- Termclustering
- Citation Clustering
- Dokumentenclustering

Die Idee des *Termclusterings* wurde Anfang der siebziger Jahre mit dem Ziel verfolgt, das Retrieval zu verbessern (vgl. [Sparck Jones, 1971]). Das Termclustering setzt bei dem Problem an, daß Wörter synonym verwendet werden. Da Synonyme auch innerhalb eines Dokumentes zu finden sind, werden alle Dokumente auf Wortpaare untersucht, die in einem Dokument gemeinsam auftreten. Wortpaare, die in vielen Dokumenten vorkommen, bilden erste Wortcluster. Da nicht nur je zwei Wörter synonym sein können, werden die initialen Cluster nach dem folgenden Prinzip verschmolzen: (a, b) und (b, c) bilden ein Cluster, dann folgt daraus ein neues Cluster (a, b, c). Die so entstandenen Termcluster sollen den Aufbau eines Thesaurus unterstützen oder zur Erweiterung der Nachfrage dienen, um ungenaue und unvollständige Dokumentbeschreibungen oder Nachfragen, die durch Variation der Wortwahl entstehen, zu ergänzen, um dadurch wiederum die Effektivität zu erhöhen. Weitere Untersuchungen in den achtziger Jahren haben aber gezeigt, daß dieser Ansatz in praktischen Retrievalumgebungen die Performanz der Systeme nicht verbessert (vgl. [Willet, 1988]).

Die beiden anderen Ansätze unterscheiden sich vom Termclustering dadurch, daß sie statt ähnlicher Terme ähnliche Dokumente zu Clustern zusammenfassen, wobei sich nach [Ingwersen, 1992] zwei Dokumente dann ähnlich sind, wenn sich ihre Inhalte ähnlich sind. Durch das Ordnen der Dokumentenmenge soll vor allem die Effizienz erhöht werden, indem gezielt in bestimmten Clustern gesucht wird, anstatt alle Dokumente zu testen. Man differenziert beide Ansätze allerdings, weil sie auf sehr unterschiedlichen Ideen beruhen:

Dem *Citation Clustering* liegt die Idee zugrunde, daß zwei Dokumente inhaltlich zueinander in Beziehung stehen, wenn entweder das eine Dokument über die Literaturangaben auf das andere Dokument verweist, oder wenn beide Dokumente auf das gleiche dritte Dokument verweisen. Deshalb ist der Einsatz von Citation Clustering nach [Ingwersen, 1992] besonders für Dokumente aus wissenschaftlichen Gebieten, in denen sehr viel mit Verweisen auf andere Literaturstellen gearbeitet wird, geeignet. In anderen Bereichen, wie z.B. im Zeitungswesen, sind ausführliche Bibliographien allerdings nicht zu finden.

Das *Dokumentenclustering* stellt einen allgemeiner verwendbaren Ansatz dar. Die charakterisierende Darstellung der Dokumente beruht nicht auf zusätzlichen Informationen, wie Bibliographien, sondern auf den Texten selbst. Dadurch kann das Dokumentenclustering auf alle Arten von Texten angewandt werden. Die einzelnen Texte werden miteinander verglichen, und es wird nach Ähnlichkeiten in der ungeordneten Dokumentenmenge gesucht.

Das Dokumentenclustering eignet sich besonders zur Weiterverarbeitung der automatisch indexierten Artikel, um sie zu Clustern von Artikeln mit gleichen Inhalten zusammenzufassen. Da sich außerdem das Termclustering hauptsächlich auf die Gestaltung der Dokumentbeschreibung und der Nachfrage konzentriert, und das Citationclustering sich aufgrund fehlender Bibliographien nicht anwenden läßt, beschäftigt sich der folgende Abschnitt 4.3.1 mit unterschiedlichen Methoden zur Clusteranalyse nur in Bezug auf das Dokumentenclustering. In Abschnitt 4.3.2 wird dann die Clusteringkomponente des Systems näher beschrieben.

4.3.1 Dokumentenclustering

Alle Methoden der Clusteranalyse beruhen zunächst auf der Berechnung einer Matrix, in der paarweise berechnete Werte eine Beziehung der Objekte untereinander festhalten. Da die meisten in der Literatur zu findenden Maße zur Berechnung dieser Werte symmetrisch sind (d.h. $S_{i,j} = S_{j,i}$), handelt es sich bei dieser Matrix um eine obere Dreiecksmatrix S:

$$S = \begin{pmatrix} \star & S_{1,2} & S_{1,3} & S_{1,4} & \cdots & S_{1,N} \\ & \star & S_{2,3} & S_{2,4} & \cdots & S_{2,N} \\ & & \star & S_{3,4} & \cdots & S_{3,N} \\ & & & \star & \ddots & \vdots \\ & & & & \ddots & S_{N-1,N} \\ & & & & & \star \end{pmatrix}$$

mit $i = 1, .., N-1, j = 1, .., N$ und $N =$ Anzahl der Objekte.

Abschnitt 4.3.1 stellt Maße zur Berechnung der Werte vor, die im Dokumentenclustering Verwendung finden, um jeweils zwei Dokumente miteinander in Beziehung zu setzen.

Ausgehend von dieser Matrix fassen die unterschiedlichen Methoden dann die Objekte zu Clustern zusammen. Abschnitt 4.3.1 skizziert diese im Hinblick auf die Einsatzmöglichkeiten für das Dokumentenclustering.

Ähnlichkeitsmaße Beim Dokumentenclustering haben sich zwei Kategorien von Maßen durchgesetzt: *Distanzmaße* wie der Euklidische Abstand zweier Vektoren

$$D_{\vec{d_i},\vec{d_j}} = \sqrt{(\vec{d_i} - \vec{d_j})^2}$$

und *Ähnlichkeitsmaße*, wie die unten beschriebenen Koeffizienten. Beide sind sowohl auf binären Termgewichtungen als auch auf den oben beschriebenen Gewichten berechenbar. Zwar bieten Distanzmaße den Vorteil der einfachen geometrischen Interpretation, aber es hat sich gezeigt, daß es nicht sinnvoll ist, das Fehlen von gleichen Termen in zwei Dokumenten als Zeichen für Gemeinsamkeiten der beiden Dokumente zu interpretieren. Es kann nämlich vorkommen, daß zwei Dokumente, die keinen Term gemeinsam haben, trotzdem nur einen geringen Abstand zueinander haben.

Als Ähnlichkeitsmaße werden in der Literatur meist die folgenden Koeffizienten vorgeschlagen (z.B. von [Rasmussen, 1992] und [Murtagh, 1992]):

Dice Koeffizient:

$$S_{\vec{d_i},\vec{d_j}} = \frac{2 \cdot \vec{d_i} \cdot \vec{d_j}}{\vec{d_i}^{\,2} + \vec{d_j}^{\,2}}$$

Jaccard Koeffizient:

$$S_{\vec{d_i},\vec{d_j}} = \frac{\vec{d_i} \cdot \vec{d_j}}{\vec{d_i}^{\,2} + \vec{d_j}^{\,2} - \vec{d_i} \cdot \vec{d_j}}$$

Kosinus Koeffizient:

$$S_{\vec{d_i},\vec{d_j}} = \frac{\vec{d_i} \cdot \vec{d_j}}{|\vec{d_i}| \cdot |\vec{d_j}|}$$

Bei binärer Termgewichtung, d.h. die Gewichte sind entweder 1 (Dokument enthält den Term) oder 0 (Dokument enthält den Term nicht), reduziert sich der Dice Koeffizient zu

$$S_{\vec{d_i},\vec{d_j}} = \frac{2C}{A + B},$$

der Jaccard Koeffizient zu

$$S_{\vec{d_i},\vec{d_j}} = \frac{C}{A + B - C}$$

und der Kosinus Koeffizient zu

$$S_{\vec{d_i},\vec{d_j}} = \frac{C}{\sqrt{A} \cdot \sqrt{B}}$$

mit A, B = Anzahl der Terme in $\vec{d_i}$ bzw. $\vec{d_j}$ und C = Anzahl der Terme die $\vec{d_i}$ und $\vec{d_j}$ gemeinsam haben.

Die drei Ähnlichkeitsmaße werden aufgrund ihrer Eigenschaften bevorzugt zum Dokumentenclustering eingesetzt: zum einen lassen sie sich leicht berechnen und zum anderen nehmen sie eine Normierung der Vektoren vor, die bei der Verarbeitung von unterschiedlich langen Texten unerläßlich ist.

Methoden Zur Clusteranalyse wurden eine Reihe von Methoden entwickelt, die unterschiedlichen Ansätzen folgen. Abbildung 3 zeigt eine zusammenfassende Unterteilung der Methoden wie sie von [Murtagh, 1992], [Everitt, 1980], [Rasmussen, 1992] und [Willet, 1988] vorgenommen wird.

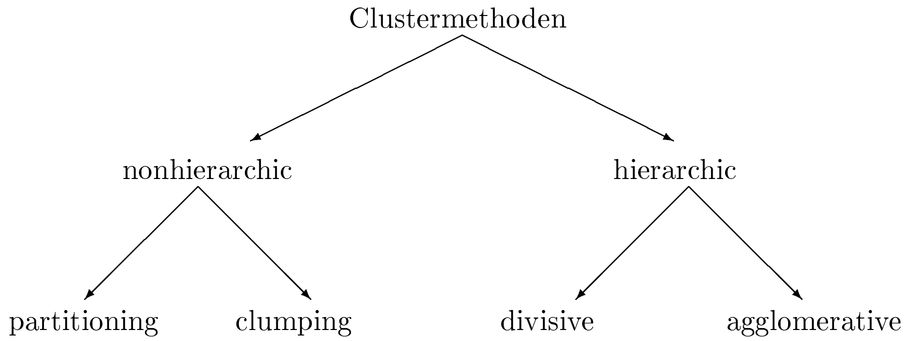

Abbildung 3: Methoden zur Clusteranalyse

Hierarchische Clustermethoden überführen die ungeordnete Menge von Objekten in eine hierarchische Ordnung. Man unterscheidet *divisive* und *agglomerative* Methoden. Dagegen teilen *nicht hierarchische* Techniken die Objektmenge in Teilmengen auf, ohne weitere mögliche Strukturen zu beachten. Zu letzteren zählen *partitioning techniques* und *clumping techniques*.

Wie bei allen Clustermethoden benötigt man auch bei den *partitioning techniques* ein Maß der Güte der gefundenen Cluster. Ziel ist es dann, alle Cluster zu finden, die dieses Gütekriterium erfüllen. Um nicht alle möglichen Aufteilungen untersuchen zu müssen, werden heuristische, a priori anzugebende Parameter, wie Anzahl und Größe der Cluster und die Lage der Clusterzentren, eingeführt. Da sich Verfahren, die in die Kategorie partitioning techniques fallen, durch das Ziel charakterisieren lassen, die vom Benutzer angegebenen Cluster hinsichtlich des Gütekriteriums zu optimieren, werden sie auch *Optimierungsverfahren* genannt.

Ausgehend von den initialen Clusterzentren werden alle Objekte einem Cluster zugewiesen, das Clusterzentrum neu berechnet und wieder alle Objekte zugewiesen bis die Berechnung der Clusterzentren keine Änderungen zur vorhergehenden Berechnung mehr aufweist.

Ebenfalls zu den partitioning techniques gehören *density search Verfahren.* Die Objekte werden hier als Punkte in einem mehrdimensionalen Raum betrachtet, dessen Dimensionen die Attribute der Objekte darstellen. Teilräume in diesem Vektorraum, die eine hohe Dichte an (Objekt-)Punkten aufweisen und die durch spärlich besiedelte Räume voneinander getrennt sind, werden als die gesuchten Cluster interpretiert.

Clumping techniques unterscheiden sich von allen anderen Clustermethoden dadurch, daß sie auch einander überlappende Cluster zulassen. Hier kann der Benutzer über Parameter Einfluß auf die Größe der Gruppen und den Grad der Überlappung nehmen.

Nicht hierarchische Verfahren werden oft zum Dokumentenclustering eingesetzt. Ihr großer Vorteil liegt im geringen Zeitaufwand zur Berechnung der Cluster, der sich allerdings nur ergibt, wenn heuristische Methoden die Suche nach einer Lösung unterstützen. Da aufgrund heuristischer Annahmen nicht alle möglichen Einteilungen in Cluster untersucht werden, muß es sich bei der gefundenen Lösung nicht um die optimale handeln. Neben der Abhängigkeit von den a priori zu bestimmenden Parametern, erweist es sich als Nachteil, daß die Verfahren auch von der Reihenfolge, in der die Dokumente eingegeben werden, abhängig sind.

Ergebnis der Clusteranalyse durch *hierarchische Verfahren* ist die Anordnung der Objekte in einem binären Baum, der auch als *Dendrogramm* bezeichnet wird. Abbildung 4 zeigt ein Beispiel.

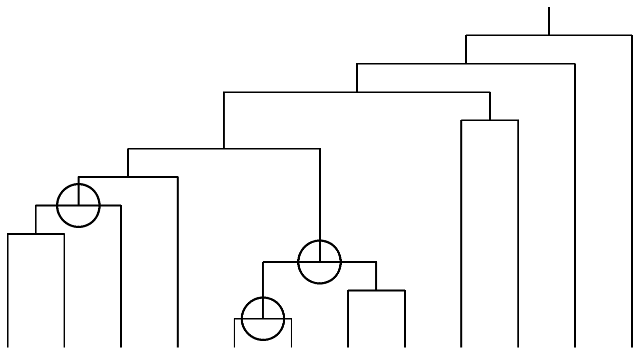

Abbildung 4: Dendrogramm

Die Blätter des Baumes repräsentieren die Objekte, während die Knoten Cluster repräsentieren. Gleichzeitig gibt jeder Knoten an, mit welcher Ähnlichkeit die Objekte des Teilbaumes, dessen Wurzel er ist, zusammengefaßt wurden. Auf jeder Ebene wird nur ein neues Cluster gebildet. Wie die Kreise in der Abbildung andeuten, entsteht ein Cluster entweder aus zwei Objekten oder zwei Clustern, oder einem bereits bestehenden Cluster wird ein weiteres Objekt hinzugefügt. Kleine Cluster auf niedrigen Ebenen weisen eine hohe Ähnlichkeit auf, wogegen die Ähnlichkeit auf höheren Ebenen, auf denen immer größere Cluster gebildet werden, abnimmt. Die Wurzel des Baumes faßt alle Objekte zu einem Cluster zusammen.

Zum hierarchischen Clustering werden *agglomerative* und *divisive* Verfahren verwendet. Divisive Verfahren gehen top-down vor, d.h. sie beginnen an der Wurzel des Baumes mit dem Cluster, das alle Objekte beinhaltet. Dieses Cluster wird nach einem Kriterium in zwei Teile geteilt, die rekursiv weiter zerteilt werden.

Je nach Wahl des Kriteriums unterscheidet man weiter *polythetische* und *monothetische* Techniken. Ist die Erfüllung des Kriteriums abhängig von mehreren Objektattributen, handelt es sich um ein polythetisches Verfahren, richtet man sich bei der Unterteilung der Cluster nur nach einem Attribut, liegt ein monothetisches Verfahren vor. Ein anschauliches Beispiel für ein monothetisches Verfahren ist die *association analysis*, die zur Clusteranalyse von Objekten mit binären Attributsausprägungen entwickelt wurde. Zur Teilung eines Clusters in zwei kleinere Cluster wird nach heuristischen Regeln ein Attribut ausgewählt, so daß alle Objekte, deren Wert für das gewählte Attribut 1 ist, zu einem Cluster gehören, und die Objekte mit der Attributsausprägung 0 das andere Cluster bilden.

Zum Dokumentenclustering sind monothetische Verfahren nicht geeignet, weil Terme, einzeln betrachtet, keinen Einfluß auf die Zuordnung eines Dokumentes zu einem

Cluster haben. Erst die Kombination von Termen, die zusammen in einem Dokument auftreten, charakterisieren das Dokument. Außerdem können auf höheren Ebenen vorgenommene Teilungen nicht wieder rückgängig gemacht werden. Dies ist ein Nachteil, den alle hierarchischen Verfahren gegenüber den nicht hierarchischen Verfahren haben. Besonders schwerwiegend ist er aber für die monothetischen Ansätze, bei denen die Gefahr der fehlerhaften Aufteilung eines Clusters aufgrund eines einzigen Attributs höher ist. Divisive, polythetische Verfahren sind in der Praxis relativ selten. Die meisten entstanden in Analogie zu den agglomerativen Verfahren (vgl. [Panyr, 1986]).

Aufgrund der Seltenheit polythetischer, divisiver Verfahren und der mangelhaften Eignung monothetischer, divisiver Verfahren, zählen die agglomerativen Verfahren beim Dokumentenclustering heute zu den populärsten.

Die existierenden agglomerativen Verfahren folgen alle dem polythetischen Ansatz und berücksichtigen zur Bestimmung der Ähnlichkeit mehrere Attribute. Im Gegensatz zu den divisiven Verfahren gehen sie bottom-up vor, d.h. ausgehend von den Objekten an den Blättern des Baumes, werden die ähnlichsten Objekte zu einem Cluster zusammengefaßt, was eine Neuberechnung der Ähnlichkeit dieses Clusters zu allen anderen Objekten erfordert. Dann werden wiederholt die ähnlichsten Objekte bzw. Cluster zusammengefaßt und die Ähnlichkeiten neu berechnet, bis an der Wurzel des Baumes nur noch ein Cluster existiert. Die agglomerativen Verfahren unterscheiden sich im wesentlichen darin voneinander, auf welcher Methode die Neuberechnung der Ähnlichkeiten zwischen dem neuen Cluster und den übrigen Objekten bzw. Clustern basiert.

Bevor im folgenden fünf verschiedene Methoden vorgestellt werden, sei hier eine allgemeine Überlegung zur Berechnung der Ähnlichkeit eingefügt: Aufgrund der drei verschiedenen Zusammenschlüsse, die in Abbildung 4 durch Umkreisungen gekennzeichnet sind, müssen Objekt–Objekt-, Objekt–Cluster- und Cluster–Cluster-Ähnlichkeiten bestimmt werden. Die Möglichkeiten zur Berechnung der Objekt–Objekt-Ähnlichkeit wurden in Abschnitt 4.3.1 beschrieben und führen zur Ähnlichkeitsmatrix, bei der die hierarchischen, agglomerativen Verfahren ansetzen. Für gebildete Cluster wird je nach zugrundeliegender Methode eine Repräsentation gewählt, die ein Cluster als neues Objekt darstellt, so daß die Objekt–Cluster- und die Cluster–Cluster-Ähnlichkeit in Analogie zur Objekt–Objekt-Ähnlichkeit ermittelt werden können. Während drei der hier angeführten Methoden direkt die Objektdarstellungen vergleichen (*single link method, complete link method* und *group average method*), geht bei den *centroid/median cluster analysis methods* und bei *Ward's method* dem Vergleich ein Schritt voraus, der eine aktuelle Clusterrepräsentation ermittelt:

Single link: Die single link Methode, die oft auch als nearest neighbour Methode bezeichnet wird, definiert die Ähnlichkeit zweier Cluster als die Ähnlichkeit derjenigen Objekte in den beiden Clustern, die sich am nächsten sind.

Complete link: Die complete link oder auch furthest neighbour Methode stellt das genaue Gegenteil zur single link Methode dar: nicht die beiden Objekte, die sich am nächsten sind, sondern die, die am weitesten voneinander entfernt sind, legen die Ähnlichkeit fest.

Group average: Hier wird die Ähnlichkeit zweier Cluster als Durchschnitt der Objekt–Objekt-Ähnlichkeiten aller Objekte der beiden Cluster ermittelt.

Centroid/Median: Die gesuchte Ähnlichkeit ergibt sich aus dem Vergleich der Clusterzentren. Wenn diese als Centroide berechnet werden, ergibt sich folgendes Problem: Das neue Zentrum zweier unterschiedlich großer Cluster, die fusioniert werden, wird näher am größeren oder sogar im größeren Cluster liegen. Damit gehen die charakteristischen Eigenschaften, die das kleinere Cluster beiträgt, verloren.

Diese Verfahrensschwäche wird durch Berechnung des Clusterzentrums als Median behoben. Man nimmt an, daß die zu fusionierenden Cluster die gleiche Größe haben, so daß das neue Zentrum in deren Mitte liegt. Alle am Zusammenschluß beteiligten Cluster haben also den gleichen charakterisierenden Einfluß.

Ward's method: Ward geht davon aus, daß jede Fusion zweier Cluster und die Darstellung des neuen Clusters als Centroid den Verlust an Information bedeutet. Deshalb faßt Ward's Methode die beiden Cluster zusammen, bei deren Zusammenschluß am wenigsten Information verloren geht, d.h. bei denen die Summe der Abweichungen der einzelnen Objektdarstellungen vom Clusterzentrum am geringsten ist (vgl. [Ward, 1963]).

Die beschriebenen Methoden können sowohl mit Ähnlichkeitsmaßen als auch mit Distanzmaßen angewandt werden. Eine Ausnahme bildet Ward's Methode, die die Abweichungen vom Clusterzentrum unter Verwendung der Euklidischen Distanz mißt, und die keine korrekten Ergebnisse liefert, wenn Ähnlichkeitskoeffizienten eingesetzt werden.

Lance und Williams veröffentlichten 1967 eine allgemeine Formel, die durch Manipulation der Parameter $\alpha_i, \alpha_j, \beta$ und γ die Berechnung neuer Distanzen (Ähnlichkeiten) nach jeder der oben angeführten Methoden erlaubt (vgl. [Lance und Williams, 1967]):

Die beiden Objekte C_i und C_j werden zu einem Cluster zusammengefaßt. Die Distanz d des neuen Clusters $C_{i,j}$ zu jedem anderen Cluster C_k ist dann gegeben durch:

$$d_{C_{i,j}C_k} = \alpha_i d_{C_iC_k} + \alpha_j d_{C_jC_k} + \beta d_{C_iC_j} + \gamma |d_{C_iC_k} - d_{C_jC_k}|$$

Die folgende Tabelle gibt Auskunft über die Werte der Parameter für die einzelnen Methoden:

single link:	$\alpha_i = \alpha_j = \frac{1}{2},\ \beta = 0,\ \gamma = -\frac{1}{2}$
complete link:	$\alpha_i = \alpha_j = \frac{1}{2},\ \beta = 0,\ \gamma = \frac{1}{2}$
group average:	$\alpha_i = \frac{n_i}{n_i+n_j},\ \alpha_j = \frac{n_j}{n_i+n_j},\ \beta = \gamma = 0$
centroid:	$\alpha_i = \frac{n_i}{n_i+n_j},\ \alpha_j = \frac{n_j}{n_i+n_j},\ \beta = -\alpha_i\alpha_j,\ \gamma = 0$
median:	$\alpha_i = \alpha_j = \frac{1}{2},\ \beta = -\frac{1}{4},\ \gamma = 0$
Ward's method:	$\alpha_i = \frac{n_i+n_k}{n_i+n_j+n_k},\ \alpha_j = \frac{n_j+n_k}{n_i+n_j+n_k},\ \beta = -\frac{n_k}{n_i+n_j+n_k},\ \gamma = 0$

n_i, n_j und n_k entsprechen der Anzahl der Objekte in C_i, C_j und C_k. Die Wahl der Parameter für Centroid-, Median- und Ward's Methode implizieren die Neuberechnung der Clusterzentren nach den Formeln:

centroid und Ward's method: $C_{i,j} = \frac{n_iC_i+n_jC_j}{n_i+n_j}$

median: $C_{i,j} = \frac{C_i+C_j}{2}$

Zur Beurteilung der verschiedenen hierarchischen, agglomerativen Methoden im Hinblick auf den Einsatz zum Dokumentenclustering ist folgendes festzuhalten (vgl. auch [Willet, 1988] und [Everitt, 1980]): Die *single link method* und die *median method* weisen eine Eigenschaft auf, die „chaining" genannt wird. Sie tendieren dazu, Cluster zusammenzufassen, die nur durch eine Kette von einzelnen Dokumenten miteinander verbunden sind. Die in Abbildung 5 als Punkte dargestellten Dokumente werden zu einem Cluster zusammengefaßt, obwohl es sich augenscheinlich um zwei Cluster handelt, die nur durch dazwischen liegende, einzelne Dokumente miteinander verbunden sind.

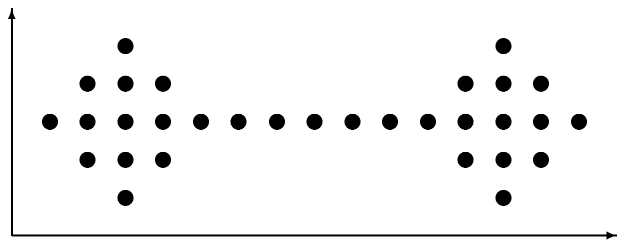

Abbildung 5: Chaining

Während also die *single link method* große, verteilte Cluster bildet, neigt die *complete link method* zur Bildung kleiner, kompakter Cluster. Unerfreuliche Effekte wie das Chaining treten nicht mehr auf. Allerdings sind auch Anwendungen denkbar, bei denen zu kleine, kompakte Cluster ebenfalls nicht wünschenswert sind. Die *group average method* bietet einen Mittelweg zwischen diesen beiden Methoden an. Alle drei Verfahren können zum Dokumentenclustering eingesetzt werden.

Die *centroid method* und die *median method* dagegen sind wegen ihrer oben angeführten Schwachstellen seltener in diesem Bereich zu finden.

Dem Verlust an Information, der besonders an der *centroid method* kritisiert wird, wirkt *Ward's method* erfolgreich entgegen. Deshalb gehört auch sie zu den populäreren Methoden. Als Nachteil erweist sich hier die Tendenz des Verfahrens, sphärische Cluster zu bilden, die den genauen Umrissen der Cluster in der gegebenen Dokumentenmenge nicht entsprechen. Da beim Dokumentenclustering bevorzugt die beschriebenen Ähnlichkeitsmaße verwendet werden, leidet das Verfahren außerdem unter der Einschränkung auf Distanzmaße.

Der folgende Abschnitt stellt die Auswahl der Verfahren für die Clusteringkomponente von AKAT vor.

4.3.2 Die Clusteringkomponente

Das Clustering folgt der automatischen Indexierung, die die berechneten Dokumentvektoren zur Weiterverarbeitung bereitstellt. Der erste Schritt im Rahmen des Clusterings ist die Berechnung der Ähnlichkeitsmatrix. Hierzu wird der Kosinuskoeffizient verwendet:

$$S_{\vec{d_i},\vec{d_j}} = \cos\alpha = \frac{\vec{d_i} \cdot \vec{d_j}}{|\vec{d_i}| \cdot |\vec{d_j}|}$$

Der Vorteil ist, daß sich der Kosinuskoeffizient leicht anschaulich verdeutlichen läßt: Es gilt, daß sich zwei Vektoren um so ähnlicher sind, je kleiner der Winkel α ist, den sie einschließen. Da die Einträge der Dokumentvektoren nur aus Gewichten $w_{i,d} \geq 0$ bestehen, können sie nur Winkel einschließen, für die $0^o \leq \alpha \leq 90^o$ gilt. Der Kosinus eines Winkels im Bereich von 0^o bis 90^o nimmt monoton ab, wenn der Winkel größer wird. Daraus folgt, daß der Kosinus des Winkels umso größer wird, je kleiner der Winkel wird, und je ähnlicher die Dokumente sich sind.

Da die Dokumentvektoren während der automatischen Indexierung bereits normiert wurden, d.h. für ihre Beträge gilt $|\vec{d_i}| = |\vec{d_j}| = 1$, reduziert sich die Berechnung der Ähnlichkeit auf das Skalarprodukt:

$$S_{\vec{d_i},\vec{d_j}} = \vec{d_i} \cdot \vec{d_j}$$

Zum Aufbau der Clusterhierarchie wird die *group average method* eingesetzt. Da es allerdings darum geht, Gruppen ähnlicher Artikel, also Artikel mit einem gemeinsamen Thema, zu extrahieren, muß der Bestimmung der Hierarchie, die der ungeordneten Dokumentenmenge zugrundeliegt, ein weiterer Verarbeitungsschritt folgen. Da das Dokumentenclustering in der Literatur als Möglichkeit zur Unterstützung des Information Retrievals beschrieben wird (vgl. [Willet, 1988]), lassen sich keine Anregungen finden, welche Cluster innerhalb der Hierarchie auszuwählen sind. Aufgrund der Dendrogramme, die bei verschiedenen Testläufen entstanden sind, lassen sich aber folgende Hypothesen aufstellen:

- Cluster, die Artikel zu einem Thema beinhalten, weisen einen höheren Ähnlichkeitswert auf als andere Cluster der Hierarchie ($\Rightarrow$ hohe Intracluster-Ähnlichkeit).

- Bei jeder Bildung eines neuen Clusters nimmt der Wert der Ähnlichkeit ab. Werden ähnliche Artikel zu einem Cluster hinzugefügt, ist die Wertabnahme geringer, als wenn ein Cluster durch unähnliche Artikel ergänzt wird ($\Rightarrow$ hohe Differenz der Ähnlichkeiten beim Übergang eines korrekten Clusters zum übergeordneten Cluster der Hierarchie).

Gemäß der signifikanten Intracluster-Ähnlichkeit und des Differenzwertes werden zwei Schwellwerte L und D eingeführt. Bei einem Cluster C_e, das auf der Ebene e gebildet wurde, und das die Intracluster-Ähnlichkeit $S^*_{C_e}$ und den Differenzwert $S^+_{C_e}$ aufweist, handelt es sich dann um ein Cluster ähnlicher Artikel, wenn gilt:

$$S^*_{C_e} \geq L \quad \text{und} \quad S^+_{C_e} \geq D$$

Die Intracluster-Ähnlichkeit $S^*_{C_e}$ kann aus dem Dendrogramm abgelesen werden, während der Differenzwert $S^+_{C_e}$ auf folgende Weise berechnet wird:

$$S^+_{C_e} = S^*_{C_e} - S^*_{C_{e+1}}$$

Dabei gibt $S^*_{C_{e+1}}$ den Ähnlichkeitswert an, mit dem das Cluster C_e auf der nächsten Ebene $e+1$ um weitere Dokumente ergänzt wird.

Wird innerhalb des Dendrogramms ein Knoten gefunden, für den die oben genannten Bedingungen gelten, werden die Artikel an den Blättern des Teilbaumes, dessen Wurzel der gefundene Knoten bildet, zu einer Gruppe von Artikeln mit einem gemeinsamen Thema zusammengefaßt.

In Anlehnung an das System Scatter/Gather (vgl. Abschnitt 2.2) wird für jedes Cluster eine Liste von zehn Wörtern ausgegeben, die innerhalb des Clusters am höchsten gewichtet wurden. Diese Liste vermittelt einen ersten Eindruck des Clusterthemas.

5 Experimente

Dieses Kapitel befaßt sich mit verschiedenen Fragestellungen, die im Rahmen der Experimente beantwortet werden sollen. Zunächst wird in Abschnitt 5.1 die Frage nach der Leistungsfähigkeit des Programms im Hinblick auf die verarbeitbare Größe der Eingabe beantwortet. Im Anschluß daran setzt sich Abschnitt 5.2 mit der Beurteilung der Clusterqualität auseinander. Die sich ergebenden Probleme und der hier gefundene Kompromiß zur Lösung werden vorgestellt.

Weiterhin wird der Nutzen hinterfragt, den die Berücksichtigung von Sprache bringen soll: Die morphologische Analyse mit GERTWOL erlaubt die Berücksichtigung von Sprache unter zwei Gesichtspunkten. Einerseits können bestimmte Wortarten ausgewählt werden und andererseits werden die im Text vorkommenden, flektierten Wortformen auf ihre Grundform zurückgeführt. Um den Nutzen bezüglich der Qualität und der Leistungsfähigkeit zu ermitteln, werden verschiedene Ansätze miteinander verglichen. Die Beobachtungen, die sich aus der Berücksichtigung von Sprache ergeben, sind in Abschnitt 5.3 zusammengefaßt.

Eine weitere, mögliche Verbesserung des Clusterings basiert auf der Gewichtung der einzelnen Bestandteile eines Artikels. Zu diesen Bestandteilen gehören z.B. die Schlagzeile, ein Untertitel und der Text. Ein Experiment zur Gewichtung einzelner Bestandteile und damit zur Berücksichtigung der Struktur von Zeitungsartikeln beschreibt Abschnitt 5.4.

Für die Experimente wurden Artikel der folgenden im Internet angebotenen Tageszeitungen herangezogen:

taz, die tageszeitung: `http://www.taz.de` [TAZ], 18.01. – 31.01.1997
Die Welt: `http://www.welt.de` [WELT], 18.01. – 31.01.1997
Süddeutsche Zeitung: `http://www.sueddeutsche.de` [SZ], 18./25.01.1997
Passauer Neue Presse: `http://www.vgp.de` [PNP], 18.01. – 31.01.1997
Schweriner Volkszeitung: `http://www.svz.de` [SVZ], 18.01. – 31.01.1997

5.1 Leistungsfähigkeit

Die folgenden Ausführungen beruhen auf Daten, die Testläufe mit Artikelkollektionen der oben genannten Zeitungen ergaben. Dabei wurden zwei Wochen lang (vom 18.01. bis zum 31.01.97) die Artikel eines Tages verarbeitet.

Die Laufzeit des Systems wird von zwei Faktoren beeinflußt. Die Verarbeitung aller Artikel eines Tages variiert stark in Abhängigkeit von der Anzahl der Artikel einer Kollektion und der Länge der Artikel (gemessen an der Anzahl der Wörter eines Artikels). Beispielsweise dauert die Kategorisierung einer Artikelkollektion aus umfangreichen Samstagszeitungen 210 Minuten (z.B. vom 18.01.97: 330 Artikeln mit durchschnittlich 440 Wörtern). Bei einer eher kleinen Kollektion von 216 Artikeln mit durchschnittlich 293 Wörtern (vom 23.01.97) sind es dagegen nur 44 Minuten.

Wie in Kapitel 4 dargestellt wurde, besteht AKAT aus drei Komponenten: aus der morphologischen Analyse durch GERTWOL, der automatischen Indexierung und der Clusteranalyse. Die Anteile an der Laufzeit verteilen sich folgendermaßen auf die Komponenten: In Bezug auf die Geschwindigkeit geben [Haapalainen und Majorin, 1994] an, daß GERTWOL auf einer SUN SPARCstation2 etwa 200 Wortformen pro Sekunde morphologisch analysiert, daß bedeutet für die Artikelkollektion vom 18.01.97 eine Laufzeit von zwölf Minuten. Bei der kleinen Kollektion vom 23.01.97 benötigt GERTWOL fünf Minuten. Der durchschnittliche Zeitanteil der morphologischen Analyse an der gesamten Verarbeitungszeit beträgt 9,6%. Einen noch kleineren Zeitanteil nimmt die automatische Indexierung in Anspruch. Mit Zeiten zwischen eineinhalb Minuten für die Artikelkollektion vom 23.01.97 und fünf Minuten für die Artikel vom 18.01.97 liegt der prozentuale Anteil an der Gesamtzeit bei 3,0%. Diese Zahlen verdeutlichen, daß die Clusteranalyse mit durchschnittlich 87,4% den größten Anteil an der Gesamtzeit beansprucht. Die Bearbeitungszeiten schwanken hier zwischen 37,5 und 192,8 Minuten (23. und 18.01.97). Die Clusteranalyse wird in zwei Schritten durchgeführt (vgl. Abschnitt 4.3. Zunächst wird die Ähnlichkeitsmatrix berechnet, wobei die Ähnlichkeit jedes Dokumentes mit allen anderen bestimmt wird. Die Rechenzeit ist also quadratisch in der Anzahl der Artikel, wobei die Länge der Artikel die Laufzeit ebenfalls beeinflußt. Der zweite Schritt der Clusteranalyse ist die Bestimmung des Dendrogramms aus der Ähnlichkeitsmatrix. Auch hier ist die Laufzeit quadratisch in der Anzahl der Artikel.

Das polynomielle Laufzeitverhalten verursacht lange Wartezeiten, wenn das System die Artikelkollektionen mehrerer Tage verarbeitet. Die Kategorisierung von 1046 Artikeln mit durchschnittlich 472 Wörtern pro Artikel dauert 880 Minuten (14 Stunden und 40 Minuten). Der Speicheraufwand liegt bei 117 Megabyte. Der Speicheraufwand ist auch der

Grund dafür, daß das System nicht in der Lage ist, 1500 Artikel mit durchschnittlich 465 Wörtern zu verarbeiten. Während der Berechnung der Ähnlichkeitsmatrix ist das System bei den zur Verfügung stehenden Kapazitäten gescheitert.

Die angegebenen CPU-Zeiten sind kein Maßstab für die Geschwindigkeit des Systems, weil sie von der Leistung des benutzten Rechners abhängig sind. Sie sollen aber einen Eindruck von der ungefähren Laufzeit und Leistungsfähigkeit vermitteln.

5.2 Clusterqualität

Die Bestimmung der Clusterqualität wirft verschiedene Probleme auf. Die in der Literatur recherchierten Verfahren, die in Abschnitt 5.2.1 kurz skizziert werden, sollen aufzeigen, welche Möglichkeiten zur Validierung von Clusterverfahren existieren. Gleichzeitig wird darauf hingewiesen, welche Eigenschaften Systeme und Anwendungsgebiete aufweisen müssen, um die beschriebenen Verfahren anwenden zu können. Aus diesen Einschränkungen leiten sich Probleme bei der Beurteilung der hier ermittelten Cluster ab, mit denen sich Abschnitt 5.2.2 befaßt. Außerdem wird der Lösungsweg beschrieben. Der Abschnitt 5.2.3 beschließt die Überlegungen zur Clusterqualität mit der Darstellung der Ergebnisse der Testläufe vom 23., 24. und 25.01.97.

5.2.1 Methoden zur Beurteilung

Anregungen zur Beurteilung der Clusterqualität wurden in zwei verschiedenen Bereichen recherchiert, zum einen im Bereich Information Retrieval und zum anderen im Bereich Clusteranalyse. Wie aus Kapitel 2 und den Abschnitten 4.2 und 4.3 hervorgeht, liegen beide Bereiche im Umfeld des hier beschriebenen Systems.

Zur Validierung von Clustertechniken nennen [Dubes und Jain, 1979] drei mögliche Ansatzpunkte:

1. Sicherstellung, daß die Objektmenge Cluster enthält und nicht nur aus zufällig zusammengestellten Objekten besteht, die keine Gemeinsamkeiten aufweisen,

2. Überprüfung, wie weit das Ergebnis des Clusterings die Struktur der gegebenen Objektmenge wiedergibt,

3. Validierung der ermittelten Cluster.

[Willet, 1988] überträgt diese Möglichkeiten auf das Dokumentenclustering zur Unterstützung des Information Retrievals und stellt fest, daß in diesem Bereich nur die ersten beiden genannten Ansatzpunkte verfolgt werden: Um sicherzustellen, daß Teile der Dokumentenmenge Gemeinsamkeiten aufweisen, wird für die gegebene Dokumentenmenge die Ähnlichkeitsmatrix S bestimmt, wie es in Abschnitt 4.3.1 beschrieben wird. Die paarweise berechneten Ähnlichkeiten $S_{i,j}$ werden in absteigender Reihenfolge sortiert und der Rang von $S_{i,j}$ in eine zweite Matrix $R(i,j)$ eingetragen. Liegt der Dokumentenmenge eine Struktur zugrunde, weist die Matrix R bestimmte Eigenschaften auf, die z.B. bei [Ling, 1975] und [Ling und Killough, 1976] nachlesbar sind.

Zur Überprüfung, ob das durch das Clustering erstellte Dendrogramm die Struktur der Dokumentenmenge wiedergibt, wird ein *Verzerrungsmaß* berechnet. Mit diesem wird erfaßt, wie stark die aus dem Dendrogramm ableitbaren Ähnlichkeiten zwischen den Objekten von den Angaben in der Ähnlichkeitsmatrix abweichen.

Zusätzlich zu den drei aufgeführten nennt [Willet, 1988] einen weiteren Ansatzpunkt: Die Untersuchung, ob der Einsatz einer Clusteranalyse die Performanz beim Information Retrieval erhöhen kann. Die Standardkriterien zur Evaluierung von Information Retrieval Systemen sind *recall* (R) und *precision* (P) (vgl. [Ingwersen, 1992]). Mit r = Anzahl relevanter, gefundener Dokumente, l = Anzahl gefundener Dokumente und c = Anzahl relevanter Dokumente in der Kollektion gilt:

$$P = \frac{r}{l} \quad \text{und} \quad R = \frac{r}{c}$$

Precision berechnet also den Anteil relevanter Dokumente an der Menge der gefundenen Dokumente. *Recall* dagegen soll den Anteil der gefundenen, relevanten Dokumente an der Menge aller relevanten Dokumente in der gesamten Kollektion bestimmen. [Shaw Jr et al., 1997] schlagen ein Maß vor, daß die beiden Kriterien *recall* und *precision* durch Bildung des harmonischen Mittels kombiniert:

$$F = \frac{2}{\frac{1}{R} + \frac{1}{P}}$$

Die Definition von *recall* führt bereits zu Einschränkungen bei der Evaluierung: Wird das zu testende IR-System auf eine beliebige, unbekannte Dokumentenkollektion angewandt, kann der Faktor c nicht genau angegeben werden. Deshalb werden die meisten Evaluierungen auf zu diesem Zweck entwickelten Kollektionen durchgeführt.

Im Bereich der Clusteranalyse steht der dritte von [Dubes und Jain, 1979] genannte Ansatzpunkt im Vordergrund, die Validierung der Cluster. Voraussetzung ist hier das Wissen eines Experten, der nur „korrekte" Zuordnungen von Objekten zu Clustern vornimmt. Das Maß erfaßt dann die Anzahl der Abweichungen der vom System zusammengestellten Cluster zur Expertenmeinung.

Im Rahmen dieser Arbeit soll allein die Qualität der Cluster überprüft werden. Deshalb kommen Verfahren, die auf die ersten zwei Ansatzpunkte eingehen, hier nicht in Betracht. Auch die beiden Kriterien *precision* und *recall* sind nicht anwendbar, weil das System keine Anfragen berücksichtigt, nach denen in IR-Systemen über die Relevanz von Dokumenten entschieden wird. Weitere Überlegungen müssen also in Richtung des dritten Ansatzpunktes gehen.

Die angeführten Verfahren verdeutlichen, daß ihnen zwei verschiedene Vorgehensweisen zugrundeliegen: Die Verfahren, die die Struktur der Objektmenge und die durch das Clustering entstandene Struktur überprüfen, verwenden heuristische Maße und vernachlässigen die Clustersemantik. Diese betrachten dagegen Verfahren, die die Relevanz von Dokumenten und die optimale Clustergestaltung einer Expertenmeinung überlassen. Beide Vorgehensweisen haben im Rahmen dieser Arbeit Probleme aufgeworfen, auf die der nächste Abschnitt eingeht.

5.2.2 Probleme

Im Rahmen dieser Experimentreihe soll zunächst die Qualität der vom System ermittelten Cluster überprüft werden, und im Anschluß daran sollen die Ergebnisse verschiedener

Verfahren miteinander verglichen werden. Wie im vorhergehenden Abschnitt geschildert, ist zur Validierung der Cluster eine Expertenmeinung notwendig, auf die später in diesem Abschnitt noch eingegangen wird.

Zum Vergleich der verschiedenen Verfahren sollten heuristische Kriterien herangezogen werden. Die Verfahren sollten anhand der *Intracluster-Ähnlichkeit* und der *Intercluster-Ähnlichkeit* verglichen werden. Bei der Intracluster-Ähnlichkeit handelt es sich um denselben Wert S_e^*, der bereits in Abschnitt 4.3.2 beschrieben wurde. Die Intercluster-Ähnlichkeit erfaßt, wie ähnlich sich die ermittelten Cluster untereinander sind. Eine Annäherung an sie stellt der Wert S_{e+1}^* dar, der jeweils die Ähnlichkeit zwischen zwei Clustern angibt, die auf der Ebene e gebildet werden. Der ebenfalls in Abschnitt 4.3.2 vorgestellte Differenzwert S_e^+ nimmt somit auch Bezug auf die Intercluster-Ähnlichkeit. Den beiden Kriterien liegt die Überlegung zugrunde, daß gute Verfahren korrekte Cluster mit hohen Ähnlichkeiten zusammenfassen, aber die Cluster untereinander sehr verschieden sind ($\Rightarrow$ geringe Ähnlichkeiten zwischen den Clustern). Als Testablauf war geplant, die Daten verschieden vorzuverarbeiten und dann automatische Indexierung und Clustering auf diese anzuwenden. Die Intracluster-Ähnlichkeit ist dann direkt aus dem Ergebnisdendrogramm ablesbar, während die Intercluster-Ähnlichkeit zwischen den ermittelten Clustern gemäß *group average* (vgl. Abschnitt 4.3.1) berechnet werden muß. Allerdings erwiesen sich diese Kriterien aus mehreren Gründen als unbrauchbar: Wie in Kapitel 4.1.2 dargestellt, werden zur Berechnung der Cluster möglichst optimale Verfahren und Maße eingesetzt. Wenn die Berechnung der Intra- und Intercluster-Ähnlichkeit auf den gleichen Maßen beruht, sind die ausgewählten Kriterien nicht unabhängig. Wählt man dagegen andere Maße als Grundlage zur Berechnung, können diese Maße nicht mehr optimal sein.

Die Unabhängigkeit der Kriterien ist aber entscheidend. Ein Testlauf, bei dem die Vorverarbeitung so gewählt ist, daß die ermittelten Cluster für einen Leser der Artikel keinen Sinn ergeben, zeigt, daß die Intracluster-Ähnlichkeit und die Intercluster-Ähnlichkeit trotzdem verhältnismäßig hoch bzw. niedrig sind. Die Qualität der Cluster kann damit also nicht quantifiziert und verglichen werden. Außerdem fogt daraus, daß die Parameter L und D keine Aussage über die Clusterqualität erlauben.

Problematisch am Einsatz dieser heuristischen Kriterien ist also, daß keine zusätzliche Information zur Beurteilung der Cluster herangezogen wird. Intra- und Intercluster-Ähnlichkeit beruhen auf denselben Informationen, mit denen die Cluster ermittelt werden. An dieser Stelle wird deutlich, daß eine Expertenmeinung notwendig ist, die die ergänzende Information liefert.

Üblicherweise wird beim Vergleich von Clusterverfahren eine Expertenmeinung zugrundegelegt, die Clustern zugehörige Objekte korrekt zuweist. Hier ist also eine Zusammenfassung aller Artikel jeweils zu einem gemeinsamen Thema gesucht. Bei den vorliegenden Datenmengen handelt es sich um über 200 Artikel fünf verschiedener Zeitungen, die am selben Tag erschienen sind. Daraus ergibt sich das erste Problem bei der Erstellung einer Expertenmeinung: die Datenmenge ist sehr groß und unübersichtlich. Jeder Artikel muß vollständig gelesen werden, um zugeordnet werden zu können.

Ein zweites Problem ist die Granularität der Cluster. Artikel können beispielsweise unter globalen Themen wie *Sport*, *Politik*, *Wirtschaft* oder ähnlichen zusammengefaßt werden, oder die globalen Themen werden zu spezielleren Themengruppen aufgespalten, wie z.B. im Bereich Sport eine Aufteilung nach einzelnen Sportarten vorgenommen werden könnte. Für einen Experten ist es schwierig, eine feste Clustergranularität beizubehalten. Das liegt zum einen daran, daß die Interessen des Lesers in die Beurteilung einfließen. Ein allgemein sportbegeisterter Leser bildet eher ein Cluster „Sport“, während ein Fahrradsportler, der zu dem kein Interesse für Tennis aufbringt, die Zeitungsartikel nach Sportarten trennt. Zum anderen kann eine feste Clustergranularität nicht zugesichert werden, weil sie von den Artikeln selbst abhängig ist. Dies belegt ein Themenkomplex, der am 23./24. und 25.01.97 ausführlich in den Zeitungen behandelt wurde: Insgesamt 38 Artikel wurden in diesen drei Tagen in den zu den Tests herangezogenen Zeitungen zum Thema *Steuerreform* veröffentlicht. Während es am 23.01. nur fünf Artikel waren, betrachteten siebzehn Artikel am 24.01. und sechzehn Artikel am 25.01. das Geschehen von verschiedenen Standpunkten. Dazu gehörten u.a. Berichte über den *Inhalt der Reform*, über die *Auseinandersetzungen in der Regierungskoalition* und über *Kommentare der Oppositionsparteien*. Die Aufspaltung der fünf Artikel vom 23.01., die die *Veröffentlichung der Steuerreform* am selben Tag ankündigen, erscheint nicht sinnvoll. Dagegen liegt es nahe, die Artikel vom 24. bzw. 25.01. bezüglich der spezielleren Themen zu sortieren, wie es AKAT auch macht.

Aufgrund der angeführten Probleme ist es also nicht möglich, eine korrekte Expertenmeinung und damit eine optimale Lösung festzulegen: erstens unterscheiden sich die gebildeten Cluster je nach Leserinteresse, und zweitens ist es fraglich, ob ein Leser in der Lage ist, die große Artikelmenge zu überblicken.

Damit bleibt zur Beurteilung der Cluster nur noch folgender Weg: Die in den Testläufen ermittelten Artikelcluster und das erstellte Dendrogramm werden im Hinblick auf die zugrundeliegende Artikelmenge überprüft. Dazu muß kontrolliert werden,

- ob alle möglichen Cluster gefunden wurden,
- ob die gefundenen Cluster vollständig sind, und
- ob alle Artikel eines Clusters ein gemeinsames Thema behandeln.

Das Dendrogramm vermittelt einen Überblick über die Artikelmenge und vereinfacht die Kontrolle der Cluster. Zwar kann keine eindeutige, optimale Lösung gegeben werden, aber es kann entschieden werden, ob die Artikelcluster über ein gemeinsames Thema berichten, oder ob sie falsch zusammengefaßt wurden. Im folgenden wird unter einem *sinnvollen Clustering* also eine Lösung verstanden, die die genannten drei Bedingungen erfüllt. In unsicheren Fällen bleibt die Bewertung der Cluster offen. Der folgende Abschnitt stellt die Ergebnisse der Tests vor, die die hier beschriebene Auswertung ergeben hat.

5.2.3 Ergebnisse

Die hier angeführten Ergebnisse beruhen auf den Artikelmengen vom 23., 24. und 25.01.97. Dabei handelt es sich um 216, 214 und 321 Artikel. Während der morphologischen Analyse werden alle Wortformen auf ihre Grundform zurückgeführt und die Wörter der Artikel werden auf wesentliche Wortarten reduziert. Bei der automatischen Indexierung gehen nur Substantive, Eigennamen und Adjektive in die Dokumentvektoren ein. Der im weiteren folgende Abschnitt 5.3.1 erläutert die Auswahl dieser Wortarten näher. Aufgrund der Dokumentvektoren wird dann die Ähnlichkeitsmatrix berechnet, auf der das Clustering basiert.

Datum	Artikel	vollständige Cluster	unvollständige Cluster	fehlende Cluster	falsche Cluster	unsichere Cluster
23.01.97	216	29	0	0	4	3
24.01.97	214	27	1	0	3	2
25.01.97	321	41	0	1	5	2

Tabelle 1: Bewertung der Cluster vom 23., 24. und 25.01.97

Tabelle 5.2.3 gibt einen zahlenmäßigen Überblick, wie weit die Testläufe die oben genannten Fragestellungen erfüllen: Zu den ersten beiden Fragen geben die ersten drei Spalten Auskunft. Die Spalte *vollständige Cluster* gibt die Anzahl der vollständigen Cluster an, deren Artikel über ein gemeinsames Thema berichten. Unter der Überschrift *unvollständige Cluster* ist die Anzahl der Cluster aufgeführt, denen noch weitere Artikel hinzugefügt werden müßten und die Rubrik *fehlende Cluster* zeigt an, wie viele Cluster nicht gefunden wurden. Die Spalte *falsche Cluster* gibt an, wie viele Cluster kein gemeinsames Thema behandeln. Die Spalte *unsichere Cluster* enthält die Anzahl der Cluster, die weder sicher den vollständigen, korrekten Clustern zugeordnet, noch als Cluster verworfen werden können.

Alle Cluster im einzelnen genauer zu beschreiben, erscheint nicht sinnvoll. Interessanter ist die Vorstellung der unsicheren Cluster, um ein genaueres Bild von der Bewertungsproblematik zu vermitteln. Außerdem werden einige falsch zusammengefaßte Artikel kurz skizziert, um Schwächen des Systems aufzuzeigen.

Von den gefundenen Clustern vom 23.01.97 wurden die folgenden als unsicher eingestuft: `[11,142]`, `[1,172]` und `[167, 168]`. Artikel `11` berichtet darüber, daß der Bundesgerichtshof die Revision im Fall eines niedersächsischen Agrarindustriellen verwarf. In Artikel `142` beklagt sich der Präsident des Bundesgerichtshofes über die hohe Zahl an Revisionen, die verhandelt werden müssen. Die subjektive Beschreibung des Themenschwerpunktes dieser beiden Artikel macht die Beurteilung des Clusters fraglich. Stellt man die Personen in den Vordergrund, verlieren die Artikel den Zusammenhang, geht es aber um Revisionen am Bundesgerichtshof, bilden sie ein korrektes Cluster. Noch deutlicher wird dies bei den Clustern `[1,172]` und `[167, 168]`: Die Gemeinsamkeit der Artikel `1` und `172` bildet das Thema *Abtreibung*. Allerdings wird über dieses Thema in zwei verschiedenen Zusammenhängen berichtet: Abtreibung wegen Vergewaltigung während des Krieges (`1`) und Ansichten der niederländischen Gesundheitsministerin zur Abtreibung (`172`). Die Artikel `167` und `168` sind Kurzmeldungen zu Äußerungen Bill Clintons: zum einen kündigt er eine Reform bei der Wahlkampffinanzierung an (`167`), zum anderen schlägt er massive Einschnitte in die gesetzliche Krankenversorgung vor (`168`). Die gleiche Problematik trifft auf die beiden unsicheren Cluster vom 25.01.97 zu.

Fehlerhafte Cluster werden aus verschiedenen Gründen gebildet. Teilweise könnte, ähnlich wie bei den unsicheren Clustern, die Verschiebung des Themenschwerpunktes einen Zusammenhang zwischen den Artikeln herstellen. Im Gegensatz zu den unsicheren Clustern entsteht hier allerdings der Eindruck, daß die Verschiebung nicht zulässig ist. Beispiele sind `[15,176]` vom 23.01., `[204,209]` vom 24.01. und `[253,262]` vom 25.01.97:

15: Rückgabe Hongkongs an China
176: Franz Beckenbauer in China

204: Bundesregierung will Haushaltssperre verhängen
209: Bilanzgewinn der Bundesbank gesunken

253: Tourismus finanziert Naturreservate
262: Reisemesse in Hannover eröffnet

Erst unter globalen Themen wie *China*, *Finanzen* und *Tourismus* lassen sich die Artikelpaare jeweils zu einem Cluster zusammenfassen. Dies entspricht aber nicht der Tendenz des Systems, Artikel sehr differenziert zu gruppieren. (Diese Eigenschaft wurde in Abschnitt 5.2.2 in Bezug auf den Themenkomplex *Steuerreform* beschrieben.)

Außerdem treten fehlerhafte Cluster auf, wenn Artikel über verschiedene Personen mit den gleichen Namen berichten. Das ist aber nur selten der Fall. Die betreffenden Artikel 137 und 159 vom 25.01.97 werden in Abschnitt 5.3.1 noch einmal aufgegriffen.

Eine dritte Fehlerquelle sind Artikel, die auf gemeinsamen Signalwörtern beruhen. Ein Beispiel sind die Artikel 39 und 46 vom 23.01.97. Es handelt sich um Buchbesprechungen in der Sparte Belletristik (39) und der Sparte Bilderbücher (46), die mit Preis-, Seiten- und Verlagsangabe enden. Ein Cluster zum Thema *Buchbesprechungen* ist zwar akzeptabel, aber die Signalwörter sind nicht prägnant genug, um alle Buchbesprechungen in einem Cluster zu erfassen.

Wie die einleitende Tabelle belegt, ermittelt das vorliegende System überwiegend korrekte Cluster. Zwar ist das Clustering nicht völlig fehlerfrei, aber die Anzahl der Fehler ist klein. Aufgrund einer nicht eindeutig bestimmbaren, optimalen Expertenmeinung bleiben unsichere Cluster, deren Korrektheit offen ist. Aber auch deren Anzahl ist klein im Vergleich zu den richtig bestimmten Clustern.

5.3 Berücksichtigung von Sprache

Die morphologische Analyse bietet die Möglichkeiten zur Auswahl der Wortarten und zur Reduzierung der im Text auftretenden Wortformen auf ihre Grundformen. Aus der Kombination dieser beiden Eigenschaften ergeben sich folgende Ansätze:

1. Nur bestimmte Wortarten werden bei der charakterisierenden Beschreibung der Dokumente berücksichtigt und die Wörter werden auf ihre Grundform zurückgeführt (MWMG).

2. Auch hier werden nur bestimmte Wortarten berücksichtigt, aber die Wörter werden nicht auf ihre Grundform zurückgeführt (MWOG).

3. Alle im Text vorkommenden Wörter werden in ihrer Grundform in den Dokumentvektor aufgenommen (OWMG).

4. Auf den Einsatz von GERTWOL wird verzichtet, so daß alle Wörter unverändert im Dokumentvektor zu finden sind (OWOG).

(Die Abkürzungen bedeuten: Mit/Ohne Wortartauswahl, Mit/Ohne Grundformreduktion.) Jeweils die Artikel eines Tages werden nach diesen Ansätzen vorverarbeitet, also morphologisch analysiert und automatisch indexiert, und zu Clustern zusammengefaßt. Grundlage für den Vergleich der verschiedenen Ansätze bilden die Clusterhierarchien und die Artikelcluster.

Das Programm bietet verschiedene Möglichkeiten, die Ergebnisse des Clusterings zu beeinflussen. Zunächst muß eine Auswahl der Wortarten getroffen werden, mit denen die Dokumente gefiltert werden. Die Tests zur Ermittlung einer sinnvollen Wortartauswahl und ihre Resultate werden in Abschnitt 5.3.1 vorgestellt.

Mit einer festen Wortartauswahl kann dann ein Vergleich der vier Ansätze MWMG, MWOG, OWMG und OWOG erfolgen. Dazu müssen die in Abschnitt 4.3.2 beschriebenen Parameter L und D festgelegt werden. Abschnitt 5.3.2 erläutert diese zweite Möglichkeit zur Beeinflussung der Clusteringergebnisse.

Den Begründungen zur Wortartauswahl und zur Parameterwahl schließt sich der Abschnitt 5.3.3 an. Es werden Ergebnisse dargestellt, die sich während der Experimente beobachten ließen, und die sich aus der Berücksichtigung der Sprache ergeben.

5.3.1 Wortartauswahl

In Kapitel 3 wurde auf Besonderheiten der Domäne „Zeitung“ und auf Eigenschaften der deutschen Sprache hingewiesen, die eine Filterung des Dokumentvektors auf spezielle Wortarten nahelegen. Damit soll eine Beschränkung der charakterisierenden Darstellung auf wesentliche Wörter bewirkt werden. Diese Zielsetzung wird auch mit der Konstruktion von Stoppwortlisten verfolgt. Leider kann hier keine Literaturstelle zu einer deutschen Stoppwortliste zitiert werden, aber die Übertragung englischer Stoppwörter, wie sie z.B. bei [Fox, 1992] aufgelistet werden, ins Deutsche zeigt, daß sich einige Wortarten generell in Stoppwortlisten wiederfinden. Zu diesen Wortarten gehören Artikel, Pronomen, Präpositionen, Adverben, Konjunktionen und Interjektionen. In Anlehnung an die Literatur werden diese Wortarten nicht mehr gesondert betrachtet. Variationsmöglichkeiten bleiben also noch bei der Kombination der Wortarten Substantive, Eigennamen, Adjektive und Verben. Abkürzungen, die bei der morphologischen Analyse auch erkannt werden, werden berücksichtigt, wenn sie gleichzeitig als Substantive oder Eigennamen analysiert werden. Abkürzungen wie „u.a.“ oder „etc.“ werden ausgeschlossen.

Aufgrund der Überlegungen in Kapitel 3 werden Substantive und Eigennamen immer als wesentliche Wortarten gewertet. Ein Test, bei dem alle Wortarten außer Substantiven und Eigennamen in den Dokumentvektor aufgenommen wurden, bestätigte die Dominanz dieser beiden Wortarten bei der Charakterisierung des Textinhaltes. Die Cluster, die während dieses Testlaufes gebildet wurden, empfindet ein Leser nicht als thematisch zusammengehörige Artikel.

Eigennamen kommt eine besondere Bedeutung zu. Darauf weisen Teilbäume hin, wie sie z.B. in der Clusterhierarchie vom 25.01.97 zu finden sind: die beiden Artikel `38` und `68` wurden zwar nicht zusammengefaßt, aber die Ähnlichkeit der beiden Artikel und der Differenzwert zum nächsten Knoten liegen nur knapp unter dem jeweiligen Schwellwert L bzw. D. Trotzdem kann ein Leser der Artikel keinen Zusammenhang zwischen der Wettervorhersage (`38`) und einem Bericht über Bundesliga-Gewichtheben (`68`) feststellen. Der Grund für die relativ hohe Ähnlichkeit der beiden Artikel liegt in der Ankündigung „kaum

Regen“ der Wettervorhersage und dem Vereinsnamen „TSV *Regen*“ einer Gewichtthebermannschaft.

Ein weiteres Beispiel für die Prägnanz der Eigennamen beim Clustering ist die Zusamenfassung der Artikel `137` und `159`. Während es sich bei Artikel `137` um eine Theaterkritik der „Don Carlos“-Inszenierung des Regisseurs K.D. *Schmidt* handelt, geht es in Artikel `159` um den Schriftsteller Arno *Schmidt*. Da die Artikel thematisch nicht soweit voneinander entfernt sind wie die Artikel `38` und `68` und somit noch weitere gemeinsame Wörter beinhalten, bilden sie für das Verfahren ein Cluster. In diesem Fall hat die prägnante Wirkung der Eigennamen das Verfahren zwar fehlgeleitet, aber gleichzeitig läßt dieses Beispiel auch darauf schließen, daß weniger verbreitete Namen, die in verschiedenen Artikeln auftreten, ein Indiz für die Zusammengehörigkeit dieser Artikel sind.

Ein Test zur Überprüfung der Qualität von Clustern, die nur aufgrund von Eigennamen gebildet wurden, war allerdings nicht möglich. Die morphologische Analyse durch GERTWOL erlaubt in vielen Fällen nicht, zwischen Substantiven und Eigennamen zu unterscheiden. Namen wie *Regen*, *Kohl* und *Westerwelle* zum Beispiel sind dem System nur als Substantive und nicht als Eigennamen bekannt. Ein Test kann somit nicht präzise genug zeigen, wie sich das Clustering aufgrund von Eigennamen im Vergleich zum Clustering aufgrund von Eigennamen und Substantiven verhält. Die folgende Überlegung soll aber verdeutlichen, daß die Ergänzung der Wortartenauswahl um Substantive sinnvoll ist: Zeitungen berichten nicht zwangsläufig über Personen oder Orte, sondern behandeln auch Sachthemen, die nicht mit Personen in Zusammenhang stehen. Charakteristisch für solche Themen sind Substantive, die zu dem jeweiligen Sachbereich gehören.

Die Auswahl der Wortarten wurde testweise noch durch Adjektive und Verben ergänzt. Dabei ergab sich, daß die entstandenen Cluster nicht stark voneinander abweichen, was wiederum zeigt, daß mit Substantiven und Eigennamen bereits wichtige Wortarten ausgewählt wurden. Allerdings konnte das Ergebnis des Clusterings durch Hinzunahme von Adjektiven noch geringfügig verbessert werden: Bei den Artikeln vom 24.01.97 wurden zwei Cluster jeweils um einen Artikel ergänzt (das Cluster `[67,214,88,157]` zum Thema *Eiskunstlauf-EM* um Artikel `213` und das Cluster `[78,127]` zum Thema *Generalstreik in Griechenland* um den Artikel `22`). Außerdem wurde bei dieser Artikelmenge wie auch bei den Artikeln vom 25.01.97 der falsche Zusammenschluß von Artikeln korrigiert. Am 24.01.97 betrifft diese Korrektur die Artikel `131` und `155`: Artikel `131` berichtet über das Asylgesuch der äthiopischen Fußball-Nationalmannschaft in Italien, während Artikel `155` ein Gespräch mit einem spanischen Fußballgewerkschafter zusammenfaßt. Am 25.01.97 wurde das Cluster `[214,215]` berichtigt (`214`: Indonesischer Brandstifter zu Haftstrafe verurteilt; `215`: Dealern droht in Indonesien Todesstrafe). Zwar könnte man auf den ersten Blick den Eindruck gewinnen, daß die beschriebenen Artikel eine Gemeinsamkeit aufweisen (Fußball/Indonesien), aber thematisch gehören sie nicht zusammen. Auf diese Problematik wurde auch schon in Abschnitt 5.2 hingewiesen.

Die weitere Ergänzung der Wortarten auf Adjektive, Substantive, Eigennamen und Verben veränderte die Ergebnisse des Clusterings nicht. Damit bestätigt sich die Vermutung, daß Verben keine wesentlichen Informationen zur Charakterisierung des Themas eines Zeitungsartikels beitragen. Hinzu kommt, daß sich bei den zu bearbeitenden Artikelmengen die Anzahl der Wörter um ein Viertel verringert, wenn auch die Verben aus den Texten herausgefiltert werden, was sich positiv auf die Bearbeitungszeit und den Speicheraufwand auswirkt.

Damit fällt die Wahl der Wortarten auf Eigennamen, Substantive und Adjektive. Wie bei den Versuchen, deren Ergebnisse in Abschnitt 5.3.3 geschildert werden, mußten auch bei den in diesem Abschnitt beschriebenen Tests die Parameter L und D festgelegt werden. Das Prinzip, nach dem die Bestimmung der Parameter in beiden Fällen vorgenommen wurde, wird im folgenden Abschnitt 5.3.2 erläutert.

5.3.2 Wahl der Parameter L und D

Wie bereits in Abschnitt 4.3.2 erläutert, handelt es sich bei den Parametern L und D um Schwellwerte. Überschreitet die Intracluster-Ähnlichkeit $S^*_{C_e}$ einer Gruppe von Artikeln den Wert L, gilt dies als erstes Indiz dafür, daß die Artikel ein Cluster bilden. Gleichzeitig wird die sprunghafte Abnahme der Ähnlichkeit beim Übergang von einem Knoten zum übergeordneten Knoten als Anzeichen für ein korrektes Cluster gewertet. D.h. die Differenz $S^+_{C_e} = S^*_{C_e} - S^*_{C_{e+1}}$ muß über dem Schwellwert D liegen. Während der Durchführung der Experimente haben sich die Vermutungen bestätigt, die der Einführung der beiden Parameter zugrundeliegen: Cluster, die Artikel zu einem Thema umfassen, haben eine hohe Intracluster-Ähnlichkeit und grenzen sich von anderen Artikeln, mit denen sie auf der nächst höheren Ebene zusammengefaßt werden, durch eine hohe Differenz der zugehörigen Ähnlichkeiten ab. Abbildung 6 zeigt zwei Ausschnitte aus einer Clusterhierarchie vom 23.01.1997, die dieses belegen:

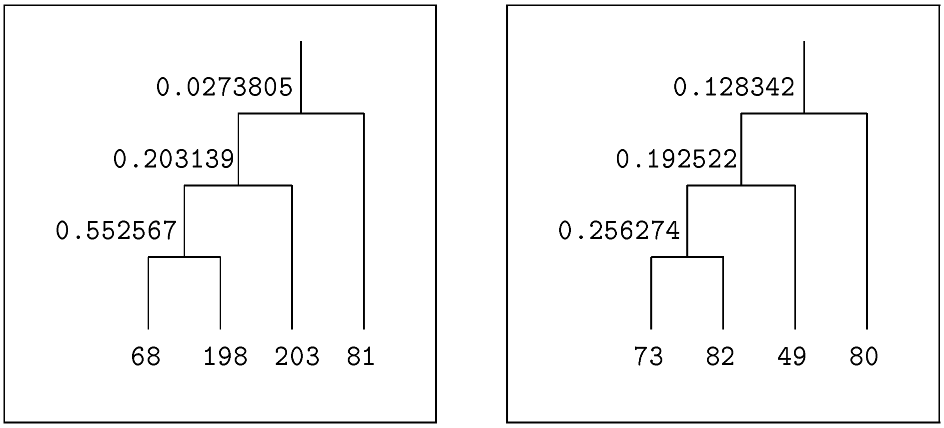

Abbildung 6: Ausschnitte aus der Clusterhierarchie vom 23.01.97

Der linke Ausschnitt zeigt das Cluster `[203,68,198]` mit der Intracluster-Ähnlichkeit $S^*_{C_e} = 0,203139$. Der hohe Betrag von $S^*_{C_e}$ und der hohe Differenzwert $S^+_{C_e} = 0,175785$ weisen darauf hin, daß die Artikel dasselbe Thema behandeln. Tatsächlich befassen sich alle drei Artikel mit dem Thema *Australian Open*. Artikel `81` dagegen kündigt eine Programmänderung bei der Weltcup-Abfahrt der Damen in Cortina d'Ampezzo an.

Der rechte Ausschnitt weist ebenfalls hohe Ähnlichkeiten $S^*_{C_e}$ auf, die Differenzen $S^+_{C_e}$ sind mit Beträgen um 0.06 aber vergleichsweise niedrig. Die Themen der Artikel sind *Eishockeyspieltermine* (`73`), *Sport im Fernsehen* (`82`), *Kulturkalender* (`49`) und *Fußballtagungstermine* (`80`).

Die niedrigen Differenzwerte $S^+_{C_e}$ bei Teilbäumen, die mit dem rechts abgebildeten Teilbaum vergleichbar sind, lassen sich folgendermaßen begründen: hohe Ähnlichkeiten bei Artikeln beruhen auf gemeinsamen Wörtern. Man kann zwei Arten von Wörtern differenzieren: zum einen Wörter, die bereits gewisse Themen implizieren und oft im Zusammenhang mit anderen Wörtern der gleichen Art vorkommen. Ein Beispiel für ein solches Wort ist der Name „Cindy“, der mit dem Thema *Rinderkrankheit BSE* eng verknüpft ist und in Zusammenhang mit Wörtern wie „Großbritannien“ oder „Galloway“ fällt.

Davon unterscheiden sich zum anderen Wörter, die zwar häufig in einzelnen Artikeln vorkommen können, aber kein spezielles Thema andeuten und nicht zwangsläufig mit anderen Wörtern zusammen genannt werden. Zu dieser zweiten Kategorie gehört z.B. das Wort „Uhr“, das u.a. bei den oben angeführten Artikeln [`73,82,49,80`] für die hohe Ähnlichkeit sorgt. Diese Wörter kommen meist noch in weiteren Artikeln vor, weil sie in verschiedenen Zusammenhängen benutzt werden.

Artikel, deren Ähnlichkeit aufgrund des gemeinsamen Vorkommens von Wörtern der zweiten Kategorie hoch ist, werden in der Clusterhierarchie mit anderen Artikeln zusammengruppiert, die dieselben Wörter in anderen Zusammenhängen beinhalten. Dadurch nimmt die Ähnlichkeit auf den unterschiedlichen Hierarchieebenen langsamer ab, als bei Artikeln, deren signifikante Wörter nur innerhalb des Clusters auftreten.

Damit erklärt sich auch, warum die Parameterwerte L und D an die unterschiedlichen Vorverarbeitungsansätze (MWMG, MWOG, OWMG und OWOG) angepaßt werden müssen. Dokumentvektoren, die auf ungefilterten Texten beruhen, enthalten viele Wörter der zweiten Kategorie, so daß die Differenzen geringer sind. Durch Erniedrigen des Parameters D kann das Clustering an die veränderte Vorverarbeitung angepaßt werden. Ebenso muß L der Vorverarbeitung entsprechend gesetzt werden.

Wie bereits in Abschnitt 5.2 erläutert, sagen die Werte $S^*_{C_e}$ und $S^+_{C_e}$ nichts über die Qualität der Cluster aus. Um einen Vergleich der verschiedenen Ansätze zu ermöglichen, können (und müssen) die Parameter L und D also auf die Vorverarbeitung abgestimmt werden. An dieser Stelle sind die Parameter die einzige Möglichkeit, das Clustering zu beeinflussen. Wenn sie so gewählt werden, daß die Abweichungen von den *sinnvollen* Clustern so gering wie möglich sind, können die resultierenden Cluster in Bezug auf ihre Qualität miteinander verglichen werden. In Anlehnung an Abschnitt 5.2 umfaßt ein *sinnvolles Clustering* alle sicheren Cluster und schließt alle falschen Cluster aus. Unsichere Cluster werden bei der Parameterwahl nicht berücksichtigt. Beim späteren Einsatz des Verfahrens können die Parameter dann konstant gesetzt werden, weil das Clustering immer auf der gleichen Vorverarbeitung basiert.

5.3.3 Beobachtungen

Die Experimente, die unter dem Thema Berücksichtigung von Sprache durchgeführt wurden, ergaben drei bemerkenswerte Aspekte, welchen die folgenden Abschnitte gewidmet sind. Zunächst werden in 5.3.3 die Ergebnisse des Vergleichs der vier oben genannten Ansätze geschildert. In Abschnitt 5.3.3 folgen Überlegungen, in wie weit sich die Performanz des Systems durch die Berücksichtigung von Sprache verbessert. Die Wortlisten, die das System zur Beschreibung der Cluster ausgibt (vgl. 4.3.2), erlauben Rückschlüsse auf die Auswirkungen der hier eingesetzten Technik des Natural Language Processing. Abschnitt 5.3.3 führt eine exemplarische Auswahl dieser Beschreibungen auf und erläutert, was sich aus ihnen schließen läßt.

Ergebnisse des Vergleichs In der Einleitung zu Abschnitt 5.3 werden vier Ansätze beschrieben, mit denen der Nutzen der morphologischen Analyse untersucht werden soll. Die Ergebnisse der zwölf Testläufe mit den Zeitungsartikeln vom 23., 24. und 25.01.97 und den vier Ansätzen MWMG, MWOG, OWMG und OWOG faßt Tabelle 2 zusammen. Jeweils die Ergebnisse der vier Methoden, die auf die Artikel eines Tages angewandt werden, werden miteinander verglichen. Aufgeführt werden die Übereinstimmungen mit einer sinnvollen Lösung und die Abweichungen von dieser Lösung. Dabei können Übereinstimmungen (Abweichungen) zweifach entstehen: richtige Cluster werden gebildet (nicht gebildet) und falsche Cluster werden nicht gebildet (gebildet). „Richtige Cluster“ bezeichnet hier alle Cluster, deren Artikel über ein gemeinsames Thema berichten.

Zu jedem Datum existiert eine sinnvolle Lösung. Die Festlegung dieser einen sinnvollen Lösung ist möglich, weil alle vier Methoden überwiegend gleiche (richtige und falsche) Cluster bilden. Auch die Unterschiede bei den Ergebnissen assoziieren keine alternativen, sinnvollen Lösungen. Wie oben bereits dargelegt, umfaßt eine sinnvolle Lösung alle richtigen Cluster und schließt alle falschen aus. Die Spalte *falsche Cluster* $\neg$ *gebildet* bezieht sich nicht auf alle möglichen falschen Cluster, sondern nur auf die falschen Cluster, die durch mindestens einen der vier Ansätze entstanden sind.

In Anlehnung an die Tabelle 5.2.3 wird die Spalte *Unsichere Cluster* beibehalten. Diese Cluster werden zum bewertenden Vergleich der Methoden nicht hinzugezogen, weil ihre Korrektheit offen ist (zu den Inhalten dieser Cluster vgl. Abschnitt 5.2.3).

Der Vergleich läßt folgendes erkennen: Alle vier Methoden bilden überwiegend sinnvolle Cluster. Falsche Cluster können allerdings nicht ausgeschlossen werden. Auf mögliche Fehlerquellen weist Abschnitt 5.2.3 hin. Trotzdem zeigt sich eine steigende Fehlerrate, wenn die Nutzung der Information, die die morphologische Analyse liefert, eingeschränkt wird. Die besten Ergebnisse, d.h. die größte Übereinstimmung mit der sinnvollen Lösung und die geringste Abweichung, zeigt die Kombination von Wortartauswahl und Grundformreduktion (MWMG). Die Qualität der Ergebnisse der beiden Ansätze MWOG und OWMG (mit Wortartauswahl, ohne Grundformreduktion und umgekehrt) scheint von der Beschaffenheit der Artikel abhängig zu sein. Das Clustering der Artikelmenge vom 23.01. nach MWOG weist weniger Fehler auf als das Clustering nach OWMG. Bei den Artikeln vom 24.01. ergibt sich das Gegenteil und die Fehlerrate am 25.01. stimmt bei beiden Ansätzen überein. Das Clustering ohne Nutzung morphologischer Information (OWOG) zeigt über allen drei Artikelmengen die kleinste Übereinstimmung und die größte Abweichung von der jeweiligen, sinnvollen Lösung.

Bevor im einzelnen auf verschiedene Cluster eingegangen wird, soll hier kurz noch einmal daran erinnert werden, daß eine Festlegung optimaler Cluster nicht möglich ist. In allen Fällen kann über die hier gewählten, sinnvollen Cluster diskutiert werden. Deshalb wird im folgenden zu den Clustern und besonders erwähnten Artikeln eine kurze Inhaltsangabe gegeben.

Beim Vergleich der Methoden bezüglich der Unterschiede bei der Bildung einzelner Cluster läßt sich beobachten, daß MWOG und OWOG in mehreren Fällen die gleichen unvollständigen Cluster bilden bzw. richtige Cluster nicht finden. Beispiele hierfür sind die Cluster `[2,3,4,124]` vom 23.01., `[67,88,157,213,214]` vom 24.01. und `[121,240]` vom 25.01.97. Das erste Cluster umfaßt Artikel, die über die *Erpressung der DB* zu dieser

Testläufe		Übereinstimmung mit sinnvoller Lösung			Abweichung von sinnvoller Lösung			Unsichere Cluster	
Datum	Vorverarbeitung	richtige C. gebildet	falsche C. ¬ gebildet	$\sum$	richtige C. ¬ gebildet	falsche C. gebildet	$\sum$	gebildet	¬ gebildet
23.01.	MWMG	29	2	**31**	0	4	**4**	3	0
	MWOG	25	5	**30**	4	1	**5**	2	1
	OWMG	27	1	**28**	2	5	**7**	2	1
	OWOG	24	2	**26**	5	4	**9**	2	1
24.01.	MWMG	27	3	**30**	1	2	**3**	0	2
	MWOG	23	2	**25**	5	3	**8**	1	1
	OWMG	27	0	**27**	1	5	**6**	2	0
	OWOG	24	0	**24**	4	5	**9**	1	1
25.01.	MWMG	41	2	**43**	1	5	**6**	2	0
	MWOG	40	3	**43**	2	4	**6**	1	1
	OWMG	40	3	**43**	2	4	**6**	2	0
	OWOG	34	4	**38**	8	3	**11**	1	1

Tabelle 2: Ergebnisse der Tests zum Vergleich verschiedener Möglichkeiten zur Berücksichtigung von Sprache

Zeit berichten. Artikel 4 listet dazu vorhergegangene Fälle auf, in denen die Deutsche Bahn erpresst wurde. Obwohl der Artikel zusätzliche Informationen zu diesem Thema gibt, fassen MWOG und OWOG nur die drei Artikel 2, 3 und 124 zusammen. Die Artikel 67, 88, 157, 213 und 214 befassen sich mit den Ereignissen der *Eiskunstlauf-EM in Paris*. Die Artikel 67, 88, 157 und 124 berichten über die Erfolge des Paares Wötzel/Steuer, der Artikel 67 erwähnt zusätzlich den Läufer Andrejs Vlascenko und Artikel 213 berichtet nur über diesen Läufer. Unter dem Thema Eiskunstlauf-EM erscheint die Zusammenfassung aller fünf Artikel am sinnvollsten. Trotzdem wurde der Artikel 213 von beiden Methoden ausgelassen. Das Cluster [121,240] wurde nicht gebildet, obwohl sich beide Artikel mit der *Öffnung der Wiener Philharmoniker für Musikerinnen* beschäftigen. Diese und weitere Fälle legen also den Schluß nahe, daß die Grundformreduktion das Clustering verbessert.

Die Methoden OWMG und OWOG (beide ohne Wortartauswahl) zeigen ebenfalls ähnliches, falsches Verhalten: hier fehlen Cluster ([185,187] am 23.01. und [141,143] am 24.01.) oder einem Cluster wurde ein falscher Artikel hinzugefügt ([67,70,153,230,319 + 65] am 25.01.). Bei den fehlenden Clustern handelt es sich in beiden Fällen um jeweils einen Bericht und ein Interview zu einem gemeinsamen Thema ([185,187]: *Kinofilm „Rossini“*; [141,143]: *Schauspieler Ben Becker*). Bericht und Interview unterscheiden sich in ihren Formulierungen, wobei Eigennamen und einige für das Thema charakteristische Substantive unverändert bleiben. Die Beschränkung auf diese wesentlichen Wortarten führt dazu, daß die Artikel nach MWMG und MWOG zusammengefaßt werden, während die Cluster nach OWMG und OWOG nicht erkannt werden. Andere Wörter geben den Ausschlag, die Artikel getrennt voneinander in die Clusterhierarchie einzufügen. Ebenso wird der Artikel 65 aufgrund nicht themenspezifischer Wörter dem oben genannten Cluster vom 25.01. zugeordnet. Während das Cluster [67,70,153,230,319] über den *Ski-Weltcup* berichtet, schildert Artikel 65 die Resultate des 6.Spieltages in der Deutschen Eishockey-Liga.

In einem Fall zeigen die Methoden MWMG und MWOG gemeinsames falsches Verhalten: Das Cluster [26,84,110] (*Tarifrunde 1997 für die Metall- und Elektroindustrie*) ist nach beiden Methoden unvollständig, weil der Artikel 26 fehlt. Hierbei handelt es sich um einen Artikel, in dem verschiedene kurze Meldungen zusammengefaßt wurden, von denen sich eine auch auf die Tarifrunde bezieht. Vermutlich wurden die Meldungen durch die Filterung nach Wortarten zu stark verkürzt, um einen Zusammenhang zu den beiden anderen Artikeln erkennen zu können. Die Auflistung verschiedener Meldungen ist kennzeichnend für mehrere Artikel der hier verwendeten Testsets. Zum Teil trifft eine der Kurzmeldungen wie oben beschrieben das Thema eines Clusters. Obwohl die Kurzmeldungen meist keine neuen Informationen liefern, sollen sie hier in die zugehörigen Cluster mitaufgenommen werden. Da die meisten Kurzmeldungen Themen berühren, über die kein anderer Artikel berichtet, ist es auch selten, daß sich unter den Kurzmeldungen eines Artikels zwei Meldungen finden, die zwei verschiedenen Clustern zuzuordnen sind (im Rahmen dieser Tests in einem Fall). Dann sind beide Zuordnungen akzeptabel, weil die Kurzmeldungen gleichwertig sind.

In den übrigen Fällen, in denen die Cluster von der sinnvollen Lösung abweichen, handelt es sich meist um Fehler, die nach der Methode OWOG entstanden sind. Das läßt darauf schließen, daß die Berücksichtigung von Sprache die Präzision des Dokumentenclusterings erhöht. Hinzukommt, daß die Reduktion der Dokumente auf wesentliche Wortarten den Zeit- und Speicherbedarf des Systems senkt. Abschnitt 5.3.3 geht kurz auf diesen Aspekt ein.

Leistung Dieser Abschnitt beschreibt kurz die Ergebnisse der Auswertung einer Statistik, die die Laufzeiten des Systems unter Nutzung der morphologischen Analyse mit den Laufzeiten ohne Nutzung der morphologischen Analyse vergleicht.

Durch die Filterung nach Wortarten, die die morphologische Analyse ermöglicht, reduziert sich die Anzahl der Wörter einer Artikelkollektion um 60%. Da die Laufzeit nicht nur von der Wortanzahl sondern auch von der konstanten Artikelanzahl abhängig ist, wirkt sich diese starke Reduzierung nicht in vollem Ausmaß auf die Laufzeit aus. Diese verringert sich durchschnittlich um 30%. Die Speicherplatzersparnis durch die Filterung beträgt im Durchschnitt 15%.

Unter der Voraussetzung, daß, wie normalerweise üblich, zumindest eine Grundformreduktion vorgenommen wird, daß eine morphologische Analyse also auf jeden Fall durchgeführt werden muß, reduziert sich der Zeitaufwand sogar um 38 % bei gleichbleibendem Speicheraufwand.

Diese Zahlen zeigen, daß die Vorverarbeitung mit GERTWOL neben der Verbesserung der Clusterqualität auch einen Zeitgewinn mit sich bringt, obwohl GERTWOL selbst Zeit in Anspruch nimmt.

Clusterbeschreibungen Die folgenden Clusterbeschreibungen, die aus den zehn höchstbewerteten Wörtern innerhalb jedes Clusters bestehen, bestätigen die Überlegungen, die in den vorangegangenen Kapiteln bezüglich der Wortwahl dargestellt wurden.

23.01.97

```
A) Cluster: [203,68,198]
     [*muster,*halbfinale,*melbourne,*hingis,*ivanisevic,
     *sampras,*fernandez,*thomas,*steffi,*pete]

B) Cluster: [123,143]
     [oeffentlich,*staat,*papier,staatlich,*staatsaufgabe,
      privat,*kommission,*s*p*d,*verwaltung,*vernunft]
```

24.01.97

```
A) Cluster: [85,44,198]
     [*philharmoniker,*orchester,*frau,*wiener,*resel,
     *kunstministerium,*maennerbund,*staatsopernorchester,
     *maennerbastion,*kontinuitaet]

B) Cluster: [176,173,174,15,77]
     [*wulff,*bluem,*kohl,*waigel,*theo,*kanzler,*f*d*p,
     *steuerreform,*sitzung,*mehrwertsteuererhoehung]

C) Cluster: [119,6,179,4,3,2,80]
     [*rind,*grossbritannien,*tier,*bauer,*entschaedigung,
     britisch,*b*s*e,*herkunftsnachweis,*hof,*toetung]
```

```
D) Cluster: [11,81,12,132]
     [*jelzin,*amtsenthebung,*duma,*parlament,*praesident,
     *kreml,*kommunist,*gesundheitsgrund,russisch,*moskau]

E) Cluster: [24,76,123]
     [*algerien,*autobombe,*fundamentalist,*regime,*macht,
     fundamentalistisch,algerisch,*g*i*a,*buergerwehr,*boufarik]

F) Cluster: [170,171]
     [*holocaust,moralisch,amerikanisch,*folge,ethisch,*pattern,
     *fernsehserie,*israel,*opfer,of]
```

25.01.97

```
A) Cluster: [116,294]
     [*fujimori,*lima,japanisch,*residenz,*peru,*geisel,
     *m*r*t*a,peruanisch,*geiselnahme,*geiseldrama]

B) Cluster: [131,197]
     [*schweden,schwedisch,elektronisch,*justizsenatorin,
     *ueberwachung,*gefaengnis,*peschel-*gutzeit,*hausarrest,
     *antrag,*modellversuch]

C) Cluster: [113,308]
     [*kommission,*besteuerung,*rente,*waigel-*kommission,
     *einkunft,*prozent,*d*m,*vorschlag,steuerfrei,
     *steuerverguenstigung]
```

Das Cluster `24A` beinhaltet den Artikel, der beispielhaft in Abschnitt 3.2 vorgestellt wurde. Substantive aus diesem Artikel finden sich auch in der Clusterbeschreibung wieder. Die Beschreibungen der Cluster `23A` und `24B` verdeutlichen die entscheidende Rolle von Eigennamen. `23B`, `24E`, `24F` und `25B` weisen eine Reihe von informationstragenden Adjektiven auf, wie *öffentlich, staatlich, fundamentalistisch, algerisch, moralisch, amerikanisch* oder *elektronisch*. Das zeigt, daß auch Adjektive einen Beitrag zur charakteristischen Beschreibung von Dokumenten (oder Clustern) leisten. Einige der Substantive, die die Clusterbeschreibungen `24C`, `24D`, `25A` und `25C` aufweisen, gehören zu den typischen Erscheinungen der Pressesprache: Die Substantive *Mehrwertsteuererhöhung, Herkunftsnachweis, Tötung, Amtsenthebung, Gesundheitsgrund, Geiseldrama* und *Waigel-Kommission* sind beispielhaft für Nominalisierungen und Augenblickskomposita. Da die Clusterbeschreibungen Zusammenfassungen der Dokumentvektoren darstellen, ist es zulässig, die Aussagen über die Beschreibungen auf die Dokumentvektoren zu übertragen.

5.4 Berücksichtigung der Struktur

Die Tests zur Berücksichtigung der Struktur werden analog zu den in Abschnitt 5.3 beschriebenen Vergleichen durchgeführt. Auch hier werden die Parameter L und D für jeden

Testlauf möglichst optimal gewählt, und die Ergebnisse im Hinblick auf eine sinnvolle Lösung betrachtet.

Die on-line angebotenen Artikel setzen sich aus verschiedenen Bestandteilen zusammen: aus der Schlagzeile, einem Untertitel, einer Kopfzeile, einer kurzen Zusammenfassung und dem eigentlichen Text. (Diese Menge an Bestandteilen bildet allerdings eine Obermenge, denn die Bestandteile sind nicht unbedingt auch gegeben.) Hinzu kommen Angaben zum Ressort, zum Autor und der Name der Zeitung. Letztere werden hier nicht weiter beachtet. In den oben beschriebenen Versuchen werden die Bestandteile eines Artikels zu einem Dokument[3] zusammengefaßt, ohne daß z.B. zwischen Schlagzeile und eigentlichem Text unterschieden wird. In Anlehnung an die in Abschnitt 4.2.1 vorgestellte Idee, die Struktur einer wissenschaftlichen Arbeit zu berücksichtigen, beschäftigt sich dieser Abschnitt mit den Auswirkungen von Gewichtungen, die auf die Struktur eines Zeitungsartikels eingehen.

Die Titel wissenschaftlicher Arbeiten und die Schlagzeilen von Zeitungsartikeln unterscheiden sich in zwei Punkten. Zum einen sind Titel wissenschaftlicher Arbeiten im Vergleich zu Artikelüberschriften lang. Zum anderen sind die Titel meist so formuliert, daß sie bereits viele Schlagwörter zur Beschreibung des Themas enthalten, während Artikelschlagzeilen entweder ebenfalls informativ gestaltet sind oder Neugier durch kryptische Formulierungen wecken sollen. In Abschnitt 3.1 wurden diese beiden gegensätzlichen Gestaltungsmöglichkeiten erläutert. Welcher Stil gewählt wird, ist abhängig von der Zeitung und vom Ressort, dem der Artikel innerhalb einer Zeitung zugeordnet ist. Eine Analyse, die nur bezüglich der Schlagzeile durchgeführt wird, müßte also zwischen einzelnen Zeitungen und Ressorts unterscheiden. Hier liegt der Schwerpunkt aber auf der Frage, ob die Berücksichtigung der Artikelstruktur durch Gewichtung analog zur Berücksichtigung der Struktur wissenschaftlicher Arbeiten die Clusteranalyse verbessern kann.

Um die Idee der Gewichtung auf Zeitungsartikel zu übertragen, müssen die Unterschiede ausgeglichen werden. Dazu werden für die Experimente die drei Bestandteile Kopfzeile, Schlagzeile und Untertitel zusammengefaßt. Die Bezeichnung *Überschrift* bezieht sich im folgenden auf diese Zusammenfassung.

Zum Vergleich wird die Überschrift gegenüber der Zusammenfassung und dem Text einfach, zweifach, fünffach und zehnfach gewichtet. Um den entgegengesetzten Extremfall zu testen, wird die ganze Überschrift aus dem Dokument herausgenommen. Die Auswertung der Clusteringergebnisse und der entsprechenden Dendrogramme weist folgende, bemerkenswerte Punkte auf:

- Die geringste Abweichung von der sinnvollen Lösung weist die Analyse der einfachen Gewichtung auf: ein Cluster wird nicht gefunden. Allerdings werden auch fünf falsche Cluster gebildet.

- Die doppelte Gewichtung der Überschrift führt dazu, daß zwei Clustern Artikel fälschlicherweise hinzugefügt werden. Analog zur einfachen Gewichtung wird ein Cluster nicht entdeckt. Die Anzahl der falsch gebildeten Cluster beträgt acht.

[3]In den vorhergehenden Abschnitten werden die Bezeichnungen *Dokument* und *Text* synonym für die Zusammenfassung aller Artikelbestandteile verwendet. Hier bezeichnet *Dokument* diese Zusammenfassung, während *Text* den eigentlichen Text des Artikels meint.

- Die fünffache Gewichtung ruft die Bildung von elf falschen Clustern hervor. Das hat zur Folge, daß auch mehrere richtige Cluster nicht mehr entdeckt werden, weil aufgrund der Gewichtung Artikel aus richtigen Clustern falschen Clustern zugeordnet werden.

- Die zehnfache Gewichtung verstärkt die Wirkung der fünffachen Gewichtung noch weiter. Mit 25 unvollständigen oder nicht gefundenen Clustern und 13 falsch gebildeten Clustern weist sie die höchste Abweichung von der sinnvollen Lösung auf.

- Die Entfernung der Überschrift aus dem Dokument führt dagegen zu guten Ergebnissen. Mit drei Abweichungen von der sinnvollen Lösung und nur fünf falsch gebildeten Clustern liegt diese Variante nach der einfachen Gewichtung am nächsten an der sinnvollen Lösung.

Diese Ergebnisse legen folgende Schlußfolgerungen nahe: Die Gewichtung der Überschriften liefert keine zusätzliche Information, mit der die Qualität der Clusteranalyse verbessert wird. Stattdessen zeigt die Herausnahme der zur Überschrift zusammengefaßten Bestandteile Kopfzeile, Schlagzeile und Untertitel auch zufriedenstellende Ergebnisse der Clusteranalyse.

Durch starke Gewichtung (fünf- und zehnfach) wird die Analyse verfälscht. Einige Zeitungen führen Artikel unter verschiedenen Rubriken (keine Ressorts) auf, die innerhalb der Überschrift kenntlich gemacht werden. Zum Beispiel werden kurze Meldungen unter dem Titel „*Tagesschau*“ veröffentlicht (PNP), oder kurze Kommentare zu verschiedenen Themen mit „*Unterm Strich*“ eingeleitet (TAZ). Diese Schlagzeilen oder auch der Untertitel „*ReiseNotizen*“ der TAZ verursachen bei starker Gewichtung fehlerhaftes Clustering bezüglich dieser Titel.

Schon die zweifache Gewichtung kann dazu führen, daß Wörter der Überschrift ein zu hohes Gewicht erhalten, wodurch Cluster falsch gebildet werden. Die Wörter „Flüchtling“ bzw. „Baby“ tragen in den Clustern `[192,213,222]` bzw. `[7,195]` die höchsten Gewichte. Beide Wörter sind in den Überschriften der jeweiligen Artikel enthalten, werden aber in unterschiedlichen Zusammenhängen benutzt, so daß die Artikel keine sinnvollen Cluster bilden.

Da die einfache Gewichtung die besten Ergebnisse aufweist, scheint die Berücksichtigung der Struktur auf die hier aufgezeigte Weise nicht sinnvoll zu sein. Der Grund dafür liegt zum einen darin, daß die Wortwahl der Artikelüberschriften verschieden motiviert ist, was zu den gegensätzlichen Stilmitteln Auslassung und Ersparung führt (vgl. 3.1). Durch diese Stilmittel variiert der Informationsgehalt der Überschriften, so daß eine Gewichtung gleichzeitig informationstragende und informationslose Überschriften betont. Bei Anwendungen, die sich nur auf eine Zeitung oder spezielle Ressorts beziehen, könnte die Gewichtung der Überschrift vielleicht nutzbringend eingesetzt werden.

Zum anderen variiert die Strukturierung von Zeitungsartikeln, denn ein Artikel besteht nicht zwingend aus den oben angeführten Bestandteilen. Eine strenge Strukturierung ließe das Erscheinungsbild von Zeitungen monoton werden, was den Verlust an Unterhaltungswert zur Folge hätte. Die Variation der Strukturierung erschwert aber die Ausnutzung der Struktur für die Kategorisierung.

6 Zusammenfassung

Diese Arbeit untersucht eine Möglichkeit, die Informationsgewinnung aus Dokumenten, die das World Wide Web zur Verfügung stellt, zu unterstützen. Dazu werden verschiedene Verfahren aus dem Bereich Natural Language Processing, insbesondere die morphologische Analyse durch GERTWOL, und zur Textkategorisierung vorgestellt. Das gewählte Anwendungsgebiet – on-line im WWW angebotene Tageszeitungen – weist besondere Eigenschaften auf, die sich auf die Gestaltung des Systems ausgewirkt haben.

AKAT besteht aus drei Komponenten: aus der morphologischen Analyse, der automatischen Indexierung und der Clusteranalyse. Die morphologische Analyse ermöglicht es die Texte auf wesentliche Wörter zu verkürzen, so daß während der automatischen Indexierung präzisere, charakteristische Darstellungen der Dokumente erstellt werden können. Anhand dieser Darstellungen wird dann eine Clusteranalyse durchgeführt. Ergebnis des Systems ist eine Aufteilung der gegebenen Dokumentenkollektion in Kategorien, wobei jede Kategorie Artikel zu einem gemeinsamen Thema enthält. Die Kategorien werden durch eine Liste von zehn Wörtern beschrieben, die einem Leser einen Eindruck von dem Thema der Artikel vermitteln.

Experimente haben bestätigt, daß sich die Berücksichtigung von Sprache verbessernd auf die Kategorisierung auswirkt. Hinzu kommt, daß die Texte durch die Reduzierung auf wesentliche Wortarten so stark verkürzt werden, daß sich die Laufzeit trotz zusätzlichem Aufwand, den die morphologische Analyse erfordert, verringert. Weitere Experimente, die die Struktur der Dokumente berücksichtigen, zeigen, daß die strukturelle Merkmale im Bereich Presse nicht genutzt werden können.

Die Kategorisierung der Dokumente unterstützt die Informationsgewinnung unter zwei verschiedenen Aspekten: gemäß den Zielen des Information Retrievals werden einem Benutzer die Kategorien vorgelegt, so daß er bei der Suche nach einer bestimmten Information gezielt in den entsprechenden Kategorien nachforschen kann. Andererseits kann er aus dem präsentierten Themenspektrum Themen wählen, die ihn interessieren, und die jeweiligen Dokumente lesen. Dieser Weg zur Unterstützung der Informationsgewinnung entspricht den Zielen des Information Filterings.

Zwei Anwendungen des Systems, die auf diesen beiden Aspekten beruhen, und die im Gebiet Presse angesiedelt sind, sind denkbar. Das System kann im Rahmen eines Archivs eingesetzt werden, das Zeitungsartikel nach Themen kategorisiert und archiviert. Sucht der Benutzer Information zu einem Thema, können ihm die Artikel der entsprechenden Themenkategorie vorgestellt werden. Unter dem zweiten Aspekt kann das Verfahren zur Unterstützung eines Information Filtering-Systems eingesetzt werden. Dabei wird die konkrete Benutzeranfrage bezüglich eines Themas durch ein automatisch erstelltes Benutzerprofil ersetzt. Das System präsentiert dem Benutzer Artikel der Kategorien, für die er mittels Feedback Interesse bekundet hat. Ein System, das diese Idee verwirklicht, ist die **P**ersönliche **Z**eitung, die [Veltmann, 1997] beschreibt.

Das System nutzt Eigenschaften der Domäne *deutsche Tageszeitungen*. Diese Eigenschaften leiten sich aus den Kennzeichen der deutschen Pressesprache ab. Allerdings wird die Pressesprache in der Linguistik auch untersucht, um Strömungen und Veränderungen in der deutschen Gegenwartssprache ausfindig zu machen. Da sich die Kennzeichen der Pressesprache auf andere Bereiche, wie zum Beispiel auf wissenschaftliche Arbeiten, übertragen lassen, liegt die Vermutung nahe, daß auch das System auf anderen Gebieten eingesetzt werden kann. Die Übertragung auf andere Gebiete bietet einen Ansatzpunkt für zukünftige Überlegungen.

Die Arbeit hat gezeigt, daß die Berücksichtung von Sprache zur Kategorisierung von Dokumenten sinnvoll ist. Weiteres nutzbares Wissen bieten die anderen Teildisziplinen der Linguistik. In der Vergangenheit wurde der Rechenaufwand gescheut, den der Einsatz von NLP-Techniken auf höheren Sprachebenen mit sich bringt. Aber im Rahmen der Möglichkeiten, die der technische Fortschritt bietet, und im Hinblick auf die Notwendigkeit der Informationsgewinnung aus sehr großen Datenangeboten, die das einleitende Zitat von [Cowie und Lehnert, 1996] deutlich macht, sollte dieser Weg zur Verarbeitung textueller Daten in deutscher Sprache weiter beschritten werden.

A Implementierung

A.1 Aufbau

Um das System zu starten, müssen nacheinander zwei Skripten aufgerufen werden. Das erste Skript aktiviert GERTWOL und filtert die Ausgabe der morphologischen Analyse nach Wortarten. Dieses Skript kann nur auf einem Computer gestartet werden, auf dem die Software für GERTWOL installiert ist:

Aufruf: `mwms <eingabedatei>`

```
#!/bin/sh
perl /home/kimo-d/schewe/Perl/strip_cM.pl $1 |
tw-ger-pp | tw-ger -d -u -Z"~|\#" |
perl /home/kimo-d/schewe/Perl/strip_gertwol.pl > zeitung
```

Die `<eingabedatei>` wird in A.2 näher beschrieben. Die Ausgabe wird in eine Datei (namens `zeitung`) geschrieben. Diese Datei ist gleichzeitig die `<eingabedatei>` für das zweite Skript. Weitere Eingaben für dieses Skript sind das `<verzeichnis>`, in dem die Eingabedatei zu finden ist, und in das die Ausgabedateien geschrieben werden, und die Parameter `<L>` und `<D>`:

Aufruf: `clustering <verzeichnis> <eingabedatei> <L> <D>`

```
#!/bin/csh
cd $argv[1]
/home/kimo-d/schewe/Prolog/Listen9/artproc $argv[2] | \
/home/kimo-d/schewe/Prolog/Listen9/index | \
/home/kimo-d/schewe/Prolog/Listen9/documents | \
/home/kimo-d/schewe/Prolog/Listen9/similar | \
/home/kimo-d/schewe/Prolog/Listen9/cluster $argv[3] $argv[4]

cat dt_stat_a dt_stat_i dt_stat_d  dt_stat_s dt_stat_c >! dt_statistic
\rm dt_stat_*
```

Wie aus dem Skript ersichtlich ist, besteht der zweite Teil der Verarbeitung aus fünf Prozessen, die im folgenden kurz beschrieben werden:

`artproc` liest die `<eingabedatei>` und überführt sie in eine von Prolog verarbeitbare Struktur namens `article(Num,Doc,Zt,Res,Aut,Page)`, die für jeden Artikel der Kollektion angelegt und als Fakt in der internen Datenbank abgelegt wird. Unter `Num` wird die Nummer des Artikels festgehalten. `Doc` umfaßt alle Textbestandteile eines Artikels, also `Schlagzeile`, `Untertitel`, `Uebertitel`, `Abstract` und `Text`. Die fünf Bestandteile werden zusammengefaßt und die Wörter alphabetisch sortiert. Danach wird für jedes Wort die Termhäufigkeit (`Tf`) bestimmt. Damit ist die unter `Doc` gespeicherte Datenstruktur eine Liste, deren Einträge aus Paaren `[Wort,Tf]` bestehen. Unter `Zt`, `Res`, `Aut` und `Page` werden die Angaben zur `Zeitung`, zum `Ressort`, zum `Autor` und zur `Seite` abgelegt. Alle so gespeicherten Daten der Artikel werden dann an den nachfolgenden Prozeß `index` übergeben.

`index` ist für die Berechnung der Dokumenthäufigkeiten (`Df`) zuständig. Es wird eine Liste erstellt, die alle Wörter aller Artikel enthält. Dabei wird gezählt, in wie vielen Artikeln das entsprechende Wort zu finden ist. Die ebenfalls alphabetisch sortierte Liste (bestehend aus Einträgen `[Wort,Df]`) wird an den nächsten Prozeß `documents` weitergereicht.

`documents` berechnet aus den Angaben zur Termhäufigkeit und zur Dokumenthäufigkeit die Gewichte (`G`) der Wörter für jeden Artikel. Die Fakten bezüglich der Artikel (`article/6`) werden zu `article(Num,Newdoc,B,Zt,Res,Aut,Page)` aktualisiert, in dem unter `Newdoc` jetzt eine Liste abgelegt wird, die dreielementige Einträge der Form `[Wort,Tf,G]` enthält. Unter `B` wird der Betrag des Dokumentvektors gespeichert und weitergereicht, um die Berechnung der Ähnlichkeitsmatrix zu vereinfachen.

`similar` berechnet die Ähnlichkeitsmatrix aus den Dokumentvektoren, die in `documents` erstellt wurden. Die Matrix wird anschließend zusammen mit den `article/7`-Fakten an den Prozeß `cluster` übergeben.

`cluster` bestimmt das Dendrogramm und untersucht es nach Teilbäumen, die Artikelcluster gemäß der Parameter `<L>` und `<D>` darstellen. Im Anschluß daran werden die zehn höchstbewerteten Wörter innerhalb jedes Clusters berechnet.

Die temporären Ausgaben, die die Prozesse während eines Durchlaufs weiterreichen, werden jeweils von dem letzten Prozeß gelöscht, der auf die Daten zugreift. Um den Ablauf überprüfen zu können, werden die Zwischenergebnisse jedes Prozesses auch in Dateien geschrieben, die nicht mehr gelöscht werden. Nähere Angaben über diese Dateien liefert der folgende Abschnitt.

A.2 Eingabe und Ausgaben

Die Eingabe für das System , d.h. für das erste Skript ist eine Datei, die durch folgende Schlüsselwörter gegliedert ist:

`XXNummer`	Eintrag	`XXEnde`
`XXSchlagzeile`	Eintrag	`XXEnde`
`XXUntertitel`	Eintrag	`XXEnde`
`XXUebertitel`	Eintrag	`XXEnde`
`XXAbstract`	Eintrag	`XXEnde`
`XXText`	Eintrag	`XXEnde`
`XXZeitung`	Eintrag	`XXEnde`
`XXRessort`	Eintrag	`XXEnde`
`XXAutor`	Eintrag	`XXEnde`
`XXSeite`	Eintrag	`XXEnde`

Jeder Eintrag bezüglich `Nummer`, `Schlagzeile` usw. wird durch das Schlüsselwort `XXEnde` beendet. Einzelne Einträge können auch leer sein. Nur der Eintrag `Nummer`, der die fortlaufende Nummerierung der Artikel enthält, und mindestens einer der Einträge `Schlagzeile`, `Untertitel`, `Uebertitel`, `Abstract` und `Text` müssen vorhanden sein.

Nach Beendigung des ersten Skripts wird das zweite aufgerufen, dessen Eingaben, wie oben bereits beschrieben, ein Verzeichnis, der Name der Eingabedatei und die Parameter `L` und `D` sind. Eine direkte Ausgabe hat das System nicht. Alle Ergebnisse werden in Dateien geschrieben. Die Ergebnisdateien lauten:

`dt_index` gibt eine Übersicht über alle Wörter, die in allen Dokumenten vorkommen, jeweils unter Angabe der Dokumenthäufigkeit.

`dt_documents` enthält für jeden Text eine Liste aller Wörter, die in ihm vorkommen, jeweils unter Angabe der Termhäufigkeit und des Gewichtes.

`dt_similar` stellt die berechnete Ähnlichkeitsmatrix als Liste von Artikelpaaren und deren Ähnlichkeit dar.

`dt_cluster` führt die zusammengefaßten Cluster auf und gibt die Ähnlichkeit an, mit der die Dokumente zusammengefaßt wurden.

`dt_tree` beinhaltet das erstellte Dendrogramm.

`dt_analyse` führt noch einmal die zusammengefaßten Cluster an und listet zu jedem Cluster die zehn höchstbewerteten Wörter auf.

`dt_statistic` enthält Angaben zum Zeit- und Speicheraufwand der einzelnen Prozesse.

`dt_index`, `dt_documents`, `dt_similar` und `dt_cluster` werden innerhalb der gleichnamigen Prozesse erstellt. `dt_tree` und `dt_analyse` werden von `cluster` veranlaßt und `dt_statistic` entsteht nach Beendigung aller Prozesse.

Literatur

[Adriaans et al., 1993] Adriaans, P., Janssen, S., und Nomden, E. (1993). Effective Identification of Semantic Categories in Curriculum Text by means of Cluster Analysis. In Adriaans, P., Hrsg., *ECML 93, Workshop notes Machine Learning Techniques and Text Analysis*, Wien.

[Belkin und Croft, 1992] Belkin, N. J. und Croft, W. B. (1992). Information Filtering and Information Retrieval: Two Sides of the same Coin? *Communications of the ACM*, 35(12):29–38.

[Brill, 1994] Brill, E. (1994). Some Advances in transformation-based Part of Speech Tagging. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, Seiten 722–727, Seattle, WA.

[Bünting, 1987] Bünting, K.-D. (1987). *Einführung in die Linguistik*. Athenäum Verlag, Frankfurt am Main, 12. Auflage.

[Bußmann, 1983] Bußmann, H. (1983). *Lexikon der Sprachwissenschaft*. Alfred Kröner Verlag, Stuttgart.

[Cowie und Lehnert, 1996] Cowie, J. und Lehnert, W. (1996). Information Extraction. *Communications of the ACM*, 39(1):80–91.

[Cutting et al., 1993] Cutting, D. R., Karger, D. R., und Pedersen, J. O. (1993). Constant Interaction-time Scatter/Gather Browsing of Very Large Document Collections. In *Proceedings of the 16th Annual International ACM/SIGIR Conference*, Seiten 126–135.

[Deerwester et al., 1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., und Harshman, R. (1990). Indexing by Latent Semantic Indexing. *Journal of the American Society for Information Science*, 41(6):391–407.

[DeJong, 1982] DeJong, G. (1982). An Overview of the FRUMP System. In Lehnert, W. und Ringle, M., Hrsg., *Strategies for Natural Language Processing*, Seiten 149–176. Lawrence Erlbaum Associates, Hillsdale, NJ.

[Dubes und Jain, 1979] Dubes, R. und Jain, A. K. (1979). Validity Studies in Clustering Methodologies. *Pattern Recognition*, 11:235–254.

[Dunlop und van Rijsbergen, 1993] Dunlop, M. und van Rijsbergen, C. (1993). Hypermedia and Information Retrieval. *Information Management & Processing*, 29(3):287–298.

[Eggers, 1973] Eggers, H. (1973). *Deutsche Sprache im 20. Jahrhundert*. R.Piper & Co, München.

[Everitt, 1980] Everitt, B. (1980). *Cluster Analysis*. Halsted Press, New York, 2. Auflage.

[Fox, 1992] Fox, C. (1992). Lexical Analysis and Stoplists. In Frakes, W. B. und Baeza-Yates, R., Hrsg., *Information Retrieval - Data Structures & Algorithms*, Kapitel 7, Seiten 102–130. Prentice Hall, Englewood Cliffs, New Jersey.

[Franzel, 1996] Franzel, C. (1996). Auffinden interessanter Wertebereiche in Datenbankattributen. Diplomarbeit, Universität Dortmund.

[Fuhr, 1995] Fuhr, N. (1995). Skriptum zur Vorlesung Information Retrieval. Universität Dortmund.

[Haapalainen und Majorin, 1994] Haapalainen, M. und Majorin, A. (1994). GERTWOL: Ein System zur automatischen Wortformerkennung deutscher Wörter.

[Hahn, 1986] Hahn, U. (1986). Methoden der Volltextverarbeitung in Informationssystemen: Ein State-of-the-Art-Bericht. In Kuhlen, R., Hrsg., *Informationslinguistik*, Seiten 195–216. Max Niemeyer Verlag, Tübingen.

[Hahn und Reimer, 1986] Hahn, U. und Reimer, U. (1986). Semantic Parsing and Summarizing of Technical Texts in the TOPIC System. In Kuhlen, R., Hrsg., *Informationslinguistik*, Seiten 153–193. Max Niemeyer Verlag, Tübingen.

[Harman, 1993] Harman, D. (1993). Overview of the First Text REtrieval Conference. In *Proceedings of the 16th Annual International ACM/SIGIR Conference*, Seiten 36–48.

[Hearst et al., 1995] Hearst, M. A., Karger, D. R., und Pedersen, J. O. (1995). Scatter/Gather as a Tool for Navigation of Retrieval Results. In Burke, R., Hrsg., *Working notes of the AAAI Fall Symposium on AI Applications in Knowledge Navigation and Retrieval*, Seiten 65–71, Cambridge, MA.

[Ingwersen, 1992] Ingwersen, P. (1992). *Information Retrieval Interaction*. Taylor Graham, London.

[Jacobs und Rau, 1988] Jacobs, P. S. und Rau, L. F. (1988). Natural Language Techniques for Intelligent Information Retrieval. In *Proceedings of the Eleventh International Conference on Research and Development in Information Retrieval*, Seiten 85–99.

[Jacobs und Rau, 1990] Jacobs, P. S. und Rau, L. F. (1990). SCISOR: Extracting Information from On-line News. *Communications of the ACM*, 33(11):88–97.

[Joachims et al., 1995] Joachims, T., Mitchell, T., Freitag, D., und Armstrong, R. (1995). WebWatcher: Machine Learning and Hypertext. In Morik, K. und Herrmann, J., Hrsg., *Beiträge zum 7. Fachgruppentreffen MASCHINELLES LERNEN*, Forschungsbericht Nr. 580, Seiten 145–149. Universität Dortmund, Fachbereich Informatik.

[Jüttner und Güntzer, 1988] Jüttner, G. und Güntzer, U. (1988). *Methoden der Künstlichen Intelligenz für Information Retrieval*. K.G.Saur, München.

[Koskenniemi, 1983a] Koskenniemi, K. (1983a). Two-level Model for Morphological Analysis. In Bundy, A., Hrsg., *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, Seiten 683–685, USA.

[Koskenniemi, 1983b] Koskenniemi, K. (1983b). *Two Level Morphology: a General Computational Model for Word-form Recognition and Production*. Dissertation, University of Helsinki, Helsinki.

[Kuikka und Salminen, 1997] Kuikka, E. und Salminen, A. (1997). Two-dimensional filters for structured text. *Information Processing & Management*, 33(1):37–54.

[Lance und Williams, 1967] Lance, G. und Williams, W. (1967). A General Theory of Classifactory Sorting Strategies. 1. Hierarchical systems. *Computer Journal*, 9:373–380.

[Lang, 1995] Lang, K. (1995). NewsWeeder: Learning to Filter Netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, Seiten 331–339, Lake Tahoe, CA.

[Lewandowski, 1975] Lewandowski, T. (1975). *Linguistisches Wörterbuch*, Jgg. 2. Quelle & Meyer, Heidelberg.

[Lewandowski, 1994] Lewandowski, T. (1994). *Linguistisches Wörterbuch*, Jgg. 1-3. Quelle & Meyer, Heidelberg, 6. Auflage.

[Lewis und Sparck Jones, 1996] Lewis, D. D. und Sparck Jones, K. (1996). Natural language Processing for Information Retrieval. *Communications of the ACM*, 39(1):92–101.

[Ling, 1975] Ling, R. (1975). An Exact Probability Distribution on the Connectivity of Random Graphs. *Journal of Mathematical Psychology*, 12:90–98.

[Ling und Killough, 1976] Ling, R. und Killough, G. (1976). Probability Tables for Cluster Analysis based on a theory of Random Graphs. *Journal of the American Statistical Association*, 71:293–300.

[Lüger, 1995] Lüger, H.-H. (1995). *Pressesprache.* Max Niemeyer Verlag, Tübingen, 2. Auflage.

[Lyons, 1995] Lyons, J. (1995). *Einführung in die moderne Linguistik.* C.H.Beck Verlag, München, 8. Auflage.

[Mauldin, 1991] Mauldin, M. L. (1991). Retrieval Performance in FERRET: A Conceptual Information Retrieval System. In *Proceedings of the 14th Annual International ACM/SIGIR Conference*, Seiten 347–355.

[Miike et al., 1994] Miike, S., Itoh, E., Ono, K., und Sumita, K. (1994). A Full-Text Retrieval System with a Dynamic Abstract Generation Function. In *Proceedings of ACM SIGIR 94*, Seiten 152–161.

[Morik, 1995a] Morik, K. (1995a). Skriptum zur Vorlesung Einführung in die Künstliche Intelligenz. Universität Dortmund.

[Morik, 1995b] Morik, K. (1995b). Skriptum zur Vorlesung Natürlichsprachliche Systeme. Universität Dortmund.

[Murtagh, 1992] Murtagh, F. D. (1992). Cluster Analysis Using Proximities. In Van Mechelen, I., Hampton, J., Michalski, R. S., und Theuns, P., Hrsg., *Categories and Concepts: Theoretical Views and Inductive Data Analysis*, Kapitel 9, Seiten 225–245. Academic press, London.

[Panyr, 1986] Panyr, J. (1986). *Automatische Klassifikation und Information Retrieval.* Max Niemeyer Verlag, Tübingen.

[Raith, 1988] Raith, W. (1988). *Gut schreiben: ein Leitfaden.* Campus Verlag, Frankfurt/Main.

[Rasmussen, 1992] Rasmussen, E. (1992). Clustering Algorithms. In Frakes, W. B. und Baeza-Yates, R., Hrsg., *Information Retrieval - Data Structures & Algorithms*, Kapitel 16, Seiten 419–442. Prentice Hall, Englewood Cliffs, New Jersey.

[Rau und Jacobs, 1991] Rau, L. F. und Jacobs, P. S. (1991). Creating Segmented Databases for Free Text for Text Retrieval. In *Proceedings of the 14th Annual International ACM/SIGIR Conference*, Seiten 337–346.

[Reimer, 1992] Reimer, U. (1992). Verfahren der automatischen Indexierung. In Kuhlen, R., Hrsg., *Experimentelles und praktisches Information Retrieval*, Seiten 171–194. Universitätsverlag, Konstanz.

[Sahami et al., 1996] Sahami, M., Hearst, M., und Saund, E. (1996). Applying the Multiple Cause Mixture Model to Text Categorization.

[Salton und Buckley, 1988] Salton, G. und Buckley, C. (1988). Term-weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5):513–523.

[Saund, 1995] Saund, E. (1995). A Multiple Cause Mixture Model for Unsupervised Learning. *Neural Computation*, 7:51–71.

[Schneider, 1984] Schneider, W. (1984). *Deutsch für Profis.* Gruner+Jahr, Hamburg.

[Schneider und Esslinger, 1993] Schneider, W. und Esslinger, D. (1993). *Die Überschrift – Sachzwänge - Fallstricke - Versuchungen - Rezepte.* List Verlag, München.

[Seiffert, 1977] Seiffert, H. (1977). *Stil heute: eine Einführung in die Stilistik.* C.H.Beck'sche Verlagsbuchhandlung, München.

[Shaw Jr et al., 1997] Shaw Jr, W., Burgin, R., und Howell, P. (1997). Performance standards and evaluation in IR test collections: cluster-based retrieval models. *Information Processing & Management*, 33(1):1–14.

[Sparck Jones, 1971] Sparck Jones, K. (1971). *Automatic Keyword Classification.* Butterworth & Co Ltd., London.

[Strzalkowski, 1995] Strzalkowski, T. (1995). Natural Language Information Retrieval. *Information Processing & Management*, 31(3):397–417.

[Sundheim, 1992] Sundheim, B. M. (1992). Overview of the Fourth Message Understanding Evaluation and Conference. In *Proceedings of the Fourth Message Understanding Conference*, Seiten 3–29, San Mateo, CA.

[van Rijsbergen, 1979] van Rijsbergen, C. (1979). *Information Retrieval.* Butterworth & Co Ltd, London, 2. Auflage.

[Veltmann, 1997] Veltmann, G. (1997). Ein Multiagentensystem zur Erstellung eines persönlichen Pressespiegels. Diplomarbeit, Universität Dortmund.

[Ward, 1963] Ward, J. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):235–244.

[Wendlandt und Driscoll, 1991] Wendlandt, E. B. und Driscoll, J. R. (1991). Incorporating a Semantic Analysis into a Document Retrieval Strategy. In Bookstein, A. e. a., Hrsg., *Proceedings of the 14th ACM/SIGIR Conference*, Seiten 270–279, Chicago.

[Willet, 1988] Willet, P. (1988). Recent Trends in Hierarchic Document Clustering: a Critical Review. *Information Processing & Management*, 24(5):577–597.