Bundesminister für Forschung und Technologie
Forschungsbericht ID Information und Dokumentation

# C T X
# Ein Verfahren zur computergestützten Texterschließung

von

Harald H. Zimmermann
Edith Kroupa
Gerald Keil

unter Mitarbeit von: Frank W. Felzmann
Ursula Hahl
Manfred Jahn
Peter Rosenbeck
Dorothea Viets

Projektleitung: Prof. Dr. Harald H. Zimmermann

Universität des Saarlandes
5.5 Informationswissenschaft
66oo Saarbrücken

---

VORBEMERKUNG

In den Jahren 1977 bis 1982 wurde - zunächst an der Universität Regensburg (1977 - 1980), dann an der Universität des Saarlandes (1980 - 1982) - ein Forschungsprojekt durchgeführt mit dem Ziel, die Einsatzmöglichkeiten hoch entwickelter Techniken der automatischen Sprachdatenverarbeitung im Bereich der Information und Dokumentation (IuD) zu untersuchen.

Das Projekt umfasst insbesondere die Nutzung und Entwicklung von Verfahrensweisen zur computergestützten Texterschließung und deren Umsetzung in Information-Retrievalverfahren. Folgende Fragen standen im Mittelpunkt:

(1) Grundsätzliche Möglichkeiten der Nutzung linguistischer Verfahrensweisen im Bereich der Erschließung größerer Textmengen zu Dokumentationszwecken;

(2) Schaffung von Vergleichs- und Bewertungsmöglichkeiten hinsichtlich anderer, v.a. ähnlicher Strategien, und zwar am Beispiel einer ausreichenden Dokumentmenge aus einem Fachgebiet des Rechts (Datenschutzrecht);

(3) Realisierung eines Labormodells mit Blick auf eine produktionsorientierte Umsetzung der entwickelten Verfahrensweisen.

Zum Abschluss des Forschungsprojekts wird hiermit ein Bericht vorgelegt, der Motivation, die methodischen Grundlagen und die Ergebnisse darstellt. Der Überblick gliedert sich in zwei Teile.

Teil Ibeschäftigt sich zunächst mit den allgemeinen Grundlagen des Information Retrieval (IR) von textueller Information, insbesondere mit maschinellen bzw. maschinenunterstützten linguistischen Analyseverfahren. Hinzu kommt eine Beschreibung der angewendeten und entwickelten Verfahrensbausteine in einigen charakteristischen Funktionen, wobei das System vom Benutzer-

standpunkt aus betrachtet und damit auch auf das Verständnis potentieller Benutzer ausgerichtet wird.

Teil II bringt detailliertere Einblicke in die einzelnen Bausteine des Verfahrens zur Computergestützten Texterschließung (CTX), wobei die für eine konkrete Anwendung erforderlichen allgemeinen Aufwendungen erläutert und an einem Beispiel durchgängig veranschaulicht werden.

Während des Projektzeitraums waren als wissenschaftliche Mitarbeiter tätig:

Frank W. Felzmann (01.08.77-28.02.82)
Ursula Hahl (01.09.80-28.02.82)
Ludwig Hitzenberger (01.07.77-28.02.79)
Manfred Jahn (01.10.78-31.12.81)
Gerald Keil (01.09.81-31.12.81)
Josef Kopelent (01.07.78-31.08.80)
Waltraud Kopelent (01.07.78-28.02.82)
Edith Kroupa (01.11.78-28.02.82)
Norbert Lang (01.01.81-31.08.81)
Christine Schneider (01.07.77-30.09.79)
Georg Werckmeister (01.07.77-30.09.78)
Helmut Werner (01.12.78-28.02.82)
Marlies Werner (15.02.80-28.02.82)

Als studentische Hilfskräfte waren tätig:

R. v. Ammon, W. Brun, S. Dickens, T. Gabor, D. N. Hai, T. Klein, M. Line, S. Müller-Zantop, P. Rosenbeck, M. Stiegler, G. Tham, H. Wippey.

Herr Dr. J. Krause war als Mitarbeiter der Abteilung NDV der Universität Regensburg an der Planung des Projektkonzepts JUDO beteiligt.

Die Schreib- und Erfassungsarbeiten erledigten in Regensburg Frau Haas, Frau Zehentbauer, Frau Lill, Frau Niehel, Frau Zelichowsky und Frau Stich. In Saarbrücken wurde das Sekretariat von Frau Wagner geführt.

Inzwischen sind - im Rahmen eines weiteren Forschungsprojekts (TRANSIT) - anwendungsorientierte Entwicklungen des Systems CTX in Arbeit, die im Jahre 1984 zu einem entsprechenden Abschluss kommen sollen. Im Mittelpunkt dieses Vorhabens stehen die testweisen Anwendungen von CTX im Rahmen der Patentdokumentation (Deutsches Patentamt), der Literaturdokumentation (Wissenschaftszentrum Berlin) und der Verknüpfung mit intellektueller Indexierung (Fachinformationszentrum Werkstoffe). Hierbei werden v.a. Aufschlüsse mit Bezug auf die Übertragbarkeit des Systems auf andere Fachgebiete, den erforderlichen Aufwand an Systempflege und die allgemeine Verwendbarkeit in der Praxis erwartet. Erste Ergebnisse hierzu sollen im Herbst 1983 vorliegen.

Einleitung: Zum Verständnis des Begriffs "Automatische Indexierung" bei CTX

Das Verfahren CTX ist im wesentlichen als ein Spezialfall der Automatischen Indexierung zu verstehen.

Die Aufgabenstellung bestand zunächst darin, ein System zur automatischen Texterschließung zu entwickeln, das auf weitestgehend beliebige natürlichsprachige Texte anwendbar ist. Dies betrifft zuerst die strukturelle Ausformulierung der Texte. Kurze wie lange Sätze, kurze und lange Wörter sollten bearbeitet werden können; es sollte (letztlich) auch keine Rolle spielen, ob ein Satz grammatikalisch ausformuliert ist (also z.B. ein Verb enthält) oder ob der Text frei von Rechtschreibfehlern ist. Derartige Anforderungen an die Robustheit bzw. Flexibilität eines Systems führen zu folgenden trivialen Konsequenzen: Entweder sinkt bei sprachlich-technisch einfachen Lösungen die Qualität der Ergebnisse oder es werden inhaltlich wie technisch aufwendigere und damit kostspieligere Lösungen erforderlich.

In der "traditionellen" Automatischen Indexierung vergleichbarer Zielsetzung wurden bislang weitgehend einfache Lösungen angestrebt. In der Praxis eingesetzte Information-Retrieval-Systeme wie STAIRS (in seiner einfachen Form) oder DIRS/GRIPS sind hier beispielhaft zu nennen. Nach ihrer Verfahrensweise können sie bezüglich der Indexierungskomponente als zeichenorientierte Systeme betrachtet werden: Eine Bearbeitungseinheit ist dabei eine beliebige Kette aus Buchstabenzeichen, begrenzt von einem Zwischenraum oder Satz- bzw. Textzeichen. Operationen über den Wortformen ordnen diese zunächst zwei Mengen zu: der Menge der sog. STOP-Wörter oder der Menge der einen Text (im Information Retrieval spricht man von einem "Text-Dokument" oder kurz einem "Dokument") charakterisierenden Stichwörter. Über den Stichwörtern operieren ggf. sog. Stringverarbeitungsfunktionen, um diese Wortformen-Stichwörter auf Mengen gleicher Teilwortketten abzubilden (sog. Trunkierung). Dieses rein technische (d.h. nichtsprachliche) Verfahren soll der (Teil-)Schwierigkeit begegnen, dass bedeutungsgleiche oder -verwandte Wortformen eines Textes eine unterschiedliche Zeichenkette aufweisen (bekanntester Fall ist die sog. Endungsflexion, hinzu kommen Wortableitungen und Wortzusammensetzungen).

Eine Anwendung derartiger Verfahren beim Retrieval, d.h. der Suche mit Stichwörtern in einem Dokumentenbestand, setzt beim Benutzer z.T. erhebliche Vorüberlegungen voraus (sprachwissenschaftlich ausgedrückt: eine "intellektuelle" morphologische Analyse), zudem ist die technische Trunkierung allein nicht immer zuverlässig. Dies wirkt sich bei Sprachen wie dem Englischen - das in diesen Belangen einen gewissen Vorreiter darstellte - wegen einer relativ schwach ausgeprägten Flexion quantitativ nicht so sehr aus. In der Terminologie der Dokumentation: der RECALL, d.h. das Verhältnis der bei einer Suchanfrage gefundenen relevanten Dokumente zu allen über ein Stichwort (in allen Zeichenketten-Varianten) identifizierbaren Dokumenten sinkt nicht wesentlich (wenn auch häufig genug merklich). Bei Sprachen mit stärkerem Flexionsreichtum wie dem Deutschen und dem Russischen ist dieser zusätzlich erforderliche intellektuelle Aufwand bei der Recherche schon deutlicher.

Die Konsequenz dieser letztlich unbefriedigenden technischen Lösungsansätze ist es, Modelle und Verfahrensweisen zu entwickeln, die Zeichenketten weitestgehend gleichen "Inhalts" (v.a. im Bereich der Flexion/Derivation) automatisch einander zuordnen. Hierzu lässt sich eine Reihe von Alternativen denken. Sie orientieren sich alle mehr oder minder an den Regeln zur Wohlgeformtheit einer Zeichenkette in natürlichsprachigen Texten. Aufgrund der Erkenntnis, dass diese

Wohlgeformtheit nicht zuverlässig über bedeutungsunabhängige Regeln zur Kombination von Buchstaben bei der Wortbildung möglich ist (auch wenn z.B. nicht jede beliebige Zeichenkette aus Buchstaben des Alphabets "physiologisch" sprechbar ist), und der Erfahrung, dass wegen der Möglichkeit der Integration fremdsprachiger (Lehn-)Wörter nicht ein Regelsystem allein zugrundegelegt werden kann, dass zudem die historisch gewachsene natürliche Sprache viele Relikte heute nicht mehr regelhafter Wortbildungen aufweist, wird den diesbezüglichen Verfahren mehr oder minder ein Vokabular (entweder in Form von ad-hoc entwickelten Ausnahmelisten oder systematisch in Form von Lexika) zugrundegelegt.

Im Bereich des Information Retrieval (IR) hat diese Problematik zunächst dazu geführt, eine zusätzliche Indexierungshilfe anzubieten (z.B. das System PASSAT bei GOLEM), ein ähnliches Verfahren wurde bei STAIRS in die Retrievalkomponente integriert (Paket TLS). Mit der Einführung von Wortlisten und Lexika ist das gleichsam "wartungsfreie" Indexierungs- und Retrievalverfahren auf Zeichenkettenebene aufgegeben und die Stufe der (einzel-)wortorientierten Verarbeitung erreicht.

Hierbei ist man v.a. mit einem Phänomen natürlicher Sprachen konfrontiert, das als allgemeines zentrales Hemmnis für komplexere Verfahren der Sprachdatenverarbeitung gesehen werden muss: die große Vielfalt des Wortschatzes einer natürlichen Sprache, verbunden mit (historisch bedingten) sog. "Unregelmäßigkeiten" bereits auf der Ebene der Wortbildung. Dies ist keine neue Erkenntnis, bildete sie doch die Motivation für Kunst-Verkehrssprachen wie ESPERANTO und letztlich auch einen Anlass für die Thesaurus-Systeme in der Dokumentation. Für die Indexierung auf Wortebene brachte sie im Hinblick auf die kommerzielle Vermarktung solcher Systeme jedoch das Problem des Verhältnisses von Investitions- und Wartungsaufwand für die Entwicklung und Pflege der Wortlisten bzw. Lexika gegenüber dem Nutzen (beim Retrieval). Da zudem relativ rasch praktikable Lösungen gefordert waren, sind trotz des lexikalischen Ansatzes eher Ad-hoc-Lösungen entstanden als linguistisch (d.h. sprachlich-morphologisch) orientierte Lösungen.

Ähnliche Ansätze wurden beispielsweise bereits in der Forschung der 60-er Jahre verfolgt. Relativ systematisch (auf der Grundlage eines Stamm-Wörterbuchs) wurde dabei im SMART-System verfahren (vgl. SALTON 1971); andere Systeme, für die stellvertretend das INTREX-System erwähnt sei, arbeiten mit Flexions- und Suffixlisten (vgl. allgemein KUHLEN 1977, S. 36 ff). Die Praktikabilität dieser Verfahren (zumindest quantitativ gesehen) erscheint - wie eine Reihe von Anwendungen belegt - erwiesen. Sie können (wie besonders KUHLEN 1977 zeigt) v.a. wertvolle Hilfen i.S. der computergestützten systematischen Wörterbucharbeit selbst geben, indem statistische Operationen über große Textmengen zu potentiellen (d.i. noch intellektuell verifizierbaren) Verknüpfungen bzw. Regularitäten führen. Die Grenzen derartiger Verfahren zeigen sich in der letztlichen Unvollständigkeit der Zusammenführung von Wortformen (v.a. im Bereich der Wortableitungen), aber besonders in der (fehlenden) semantischen (d.h. bedeutungsmäßigen) Differenzierung von Wörtern bzw. Teilen von Wortzusammensetzungen und der fehlenden (sprachlichen) Zusammenführung von Mehrwortbegriffen (d.i. von Wortfolgen, die v.a. in Fachsprachen thematisch eine Einheit bilden; Kennzeichen eines Mehrwortbegriffs ist es häufig auch, dass das Einzelwort für sich allein nicht (mehr) die gleiche Bedeutung hat: KALTER KAFFEE, JURISTISCHE PERSON).

Zur Lösung des besonderen Problems der Identifikation von Mehrwortbegriffen gibt es in nahezu allen kommerziellen IR-Systemen Hilfslösungen. Während boole-sche Verknüpfungen (d.h. mengenlogische Operationen wie UND, ODER, UND NICHT, z.B. HAUS UND VERKAUF,

VERKAUF ODER VERLEIH; VERLEIH UND NICHT LEASING) i.a. dazu dienen, Dokumente zu identifizieren, die (irgendwie) Stich- oder Schlagwörter in der gewünschten Verknüpfung aufweisen, lässt sich vielfach die "Nähe" dieser Wörter zueinander (z.B. im gleichen Kapitel, im gleichen Satz, unmittelbar nebeneinander) formal sehr einfach notieren und bei der Suche entsprechend ausnutzen. Es handelt sich also um eine ähnliche technische (und nicht sinnbezogene) Funktion wie die Trunkierung, die es ausnutzt, dass mehrwortige Begriffe (z.B. JURISTISCHE PERSON, TREIBER IN DREIZUSTANDSLOGIK, METHODIK DES DEUTSCHUNTERRICHTS) auch physisch relativ "nahe" nebeneinander im Text vorkommen. Ähnlich der Trunkierung bedarf es beim Retrievalvorgang in derartig ausgerichteten Systemen entsprechender "intellektueller" Überlegungen und deren Umsetzung in Retrievalanweisungen, um Mehrwortbegriffe zu identifizieren.

Die Benutzung von Mehrwortbegriffen (wie übrigens auch der Komposita) ist ein sprachliches Mittel zur Erreichung einer besseren PRECISION, d.h. zur Reduktion des "Ballasts" (oder genauer: nichtrelevanter Dokumente) bei der Recherche. (In der klassischen Methodologie von Information und Dokumentation sind die Begriffe RECALL und PRECISION allgemeiner gefasst; dennoch treffen sie für die vorliegende eingeschränkte Verwendung zu.) Die abstandsorientierten technischen Verfahren zur Identifikation von Mehrwortbegriffen sind - ähnlich der Trunkierung - für die Praxis nützlich, sie sind zugleich allgemeiner verwendbar, insbesondere zur Präzisierung einer Suchanfrage im weiteren konzeptuellen Bereich; insofern wäre ein Vergleich allein im Hinblick auf verbesserte Verfahren der Mehrwortidentifikation (s.u.) sicherlich unzureichend.

Im Bereich der bedeutungsmäßigen Differenzierung von Stichwörtern/Begriffen bieten die kommerziellen Verfahren die erwähnten "indirekten" Möglichkeiten der Nutzung von Abstandsangaben an. Soweit nämlich ein einzelnes (mehrdeutiges) Wort mit anderen (ein- oder mehrdeutigen) Wörtern zur Präzisierung einer Suchanfrage mit der UND-Verknüpfung koordiniert wird, lässt sich fast immer (zumindest für praktische Zwecke ausreichend) eine gleichsam konzeptuelle Vereindeutigung im Sinne der begrifflichen Vorstellung des Recherchierenden erreichen. Vordergründig betrachtet ist also eine Indexierung unter Differenzierung der "lexikalisch möglichen" Bedeutungen eines Wortes (z.B. ANLAGE, BANK, ...) nicht nötig, wenn hei der Recherche auf die kotextuell aktualisierte Bedeutung (ANLAGE i.S. von PARKANLAGE, BANK i.S. von GELDINSTITUT) durch ein zusätzliches Vorhandensein (und Überprüfen) eines geeigneten Kontextwortes (z.B. SPAZIERWEG UND ANLAGE; BANK UND DARLEHEN) zugegriffen wird. Dennoch sind solche Argumente wenig systematisch begründet, ganz zu schweigen von Problemfällen, bei denen derartige technische Strategien nicht greifen (z.B. boolesches ODER, UND NICHT). Letztlich ist hier die Problemlösung auf eine "trickreiche" Benutzung einer im Prinzip dafür nicht entwickelten Funktion eines IR-Systems abgewälzt.

Die Forschungsprojekte JUDO bzw. JUDO-DS haben sich zum Ziel gestellt, für die hier angesprochenen Problemkreise sprachlich motivierte und sprachbezogene (d.h. stärker "linguistische") Lösungen im Modell (fachsprachenbezogen) zu entwickeln (Projekt JUDO - 1976-1980 -) und labormäßig zu erproben (Projekt JUDO-DS - Anwendung im Bereich Datenschutzrecht - 1980-1982). Neben der systematischen Einführung linguistischer Lösungsansätze (z.B. zur vollständigen Behandlung der Flexionsmorphologie) sollte - auch dies zur Überwindung bestehender Ansätze - das System (z.B. im lexikalischen Bereich) offen sein für spätere Stufen, z.B. zur maschinellen Sprachübersetzung.

Weitgehend ausgegrenzt werden bei der Systementwicklung mathematisch-statistische Verfahren (wie automatische Gewichtung von Stichwörtern, Term-Clustering, Dokument-Clustering). Diese Verfahren können allerdings auf den durch CTX "bereinigten" Input aufgesetzt werden und versprechen dann potentiell bessere Ergebnisse.

Als Basis des Verfahrens wurde zu diesem Zweck das in grundlagenorientierter Entwicklung befindliche "Saarbrücker Übersetzungssystem" (SUSY), insbesondere in den auf die Analyse der deutschen Sprache bezogenen lexikalischen wie algorithmischen Teilen, herangezogen. Die "Arbeitsteilung", die hierbei mit den Teilprojekten Al bzw. A2 des Sonderforschungsbereichs "Elektronische Sprachforschung" (SFB 100) praktiziert wurde, bestand im wesentlichen aus folgenden Aufgaben: Die Mitarbeiter des SFB waren (soweit erforderlich) bemüht, die Grundlagenentwicklung auf Problemlösungen (v.a. sprachanalytischer Art) auszurichten, die bei dem sprachlichen Datenmaterial des Forschungsprojekts JUDO bzw. JUDO-DS auftraten, soweit dies systematisch integrierbar war. Im Gegenzug wurde seitens der Projekte "realistisches" Material zur Erprobung der Regeln und Lexikon-Strukturen verfügbar: im Bereich der Erprobung selbst wurden die Analyseergebnisse durch Mitarbeiter der JUDO-Projekte ausgewertet. Teile des SUSY-Systems (z.B. auch dort entwickelte Lexika) sind somit Teile des CTX-Systems, dessen zusätzliche Software den sprachanalytischen Teil des SUSY-Systems gleichsam umrahmt (Inputvorbereitung, Outputverarbeitung). Im lexikalischen Bereich wurde allerdings nicht nur eine (meist textbezogene, z.T. auch systematische) Erweiterung des Lexikoninventars im Rahmen der JUDO-Projekte durchgeführt, sondern auch wesentliche Teile des Lexikon-Systems (v.a. die Entwicklung der Derivationslexika und des Thesaurus-Systems) neu konzipiert und materiell ausgefüllt.

Das auf diese Weise entstandene Gesamtsystem "Computergestützte Texterschliessung" (CTX) erbringt zum Abschluss der 2. Projektphase folgende wesentliche Funktionen:

- Flexionsformenwerden systematisch auf eine sie repräsentierende Grundform zurückgeführt. (Dabei werden auch Um- und Ablaute, Infigierungen, abgetrennte Verbzusätze behandelt);
- Wortableitungenwerden systematisch einander zugeordnet. Dies geschieht durch Relationierung der Wörter, so dass dem Benutzer die Möglichkeit bleibt, diese Funktion einzubringen oder wegzulassen);
- Wortzusammensetzungenwerden (unter intellektueller Kontrolle) lexikalisch (durch entsprechenden Relationen) mit sinntragenden Teilwörtern verknüpft;
- Mehrwortbegriffewerden (auf der Grundlage bestimmter syntaktischer Strukturen) identifiziert.

Diese sowohl den RECALL erhöhenden als auch die PRECISION (bei Bedarf) steigernden Funktionen werden in jedem Anwendungsfall von CTX benutzt (Morpho-syntaktisches System CTX-I). Bei bestimmten Anwendungen - dies ist v.a. bei Integration eines systematischen Thesaurus erforderlich - wird ein weiterer Systemteil zur semantischen Disambiguierung benötigt, der den morpho-syntaktischen Teil voraussetzt und integriert (Semantisches System CTX-II).

Damit sind für alle zuvor angesprochenen Problemfälle der Freitext-Indexierung entsprechende Funktionen entwickelt:

- Ermittlung von Grundformen und Aufbau von Wortableitungsrelationen bzw. Teilwort-

relationen statt Wortformen/Trunkierung

- Ermittlung natürlichsprachiger Mehrwortbegriffe anstelle komplizierter Abstandsfunktionen

- semantische Vereindeutigung mehrdeutiger Wörter statt boolesche UND-Verknüpfung beim Retrieval (bei CTX-Il).

Dass dies nur ein - wenn auch entscheidender - Ausschnitt der im Rahmen des Projekts erstellten Funktionen darstellt, zeigt die vorliegende Dokumentation.

*

Mit der linguistisch und sprachsystematisch ausgerichteten Bearbeitung der Flexionsmorphologie, der Derivations- und Kompositionsproblematik sowie der sprachbezogenen Erkennung von Mehrwortbegriffen wird von der wortbezogenen zur kontextuellen (zumindest phraseologischen) Sprachdatenverarbeitung übergegangen. Im Bereich der semantischen Disambiguierung wird ggf. nicht nur enzyklopädisches Wissen (z.B. kondensiert in einem Thesaurus), sondern auch der satzübergreifende Kontext von Bedeutung. Die dadurch angestrebte höhere Qualität (aber auch: die größere "Natürlichkeit" und Bequemlichkeit) beim Retrieval hat andererseits ihren Preis: mehr Aufwand in der System- und Lexikonpflege, größerer Aufwand an Rechenzeit und weiterer Speicherplatzbedarf. Dies kann an dieser Stelle nicht im Detail begründet und behandelt werden; manches "Negativum" (z.B. bezüglich des Rechenzeit- und Kodieraufwands) ist zudem bedingt durch die Labor- und Forschungssituation. In der Diskussion mit Anwendern und Experten im Bereich des (dokumentorientierten) Information Retrieval spielen derartige Fragen allerdings eine wichtige Rolle.

Die Entwicklung und der Transfer linguistischer Verfahren für bzw. in die Praxis stellt sich in diesem Zusammenhang als besonderes Phänomen dar. Wenn man sich andererseits vor Augen hält, dass bereits eine Unzahl von (Fach-)Thesauri und Klassifikationen entwickelt wurden, dass im "traditionellen" Lexikon-Bereich (wenn auch mit anderen Zielgruppen und Märkten) Millionenbeträge für die Entwicklung von (gedruckten) Lexika ausgegeben werden, so muss man sich fragen, wieso ein stärker systematischer Ansatz (wie es z.B. das CTX-System darstellt) nicht in der Datenverarbeitung.- und Informationsindustrie (d.h. z.B. auch ohne die Unterstützung der öffentlichen Hand) schon früher hatte entwickelt und realisiert werden können. Bis auf wenige Ausnahmen ist das linguistische Know-How des Sprachanalysesystems zu CTX bereits in den Schulgrammatiken zu finden (auch die Transformationsgrammatik ist bezüglich der verwendeten Elemente schon Anfang der 60-er Jahre begründet worden); das eingebrachte lexikalische Wissen ist ebenfalls weitestgehend in traditionellen Lexika "gespeichert".

Zwei Gründe könnten insbesondere maßgebend gewesen sein für eine derartig späte Entwicklung: Einerseits' ist jeder Laie - oberflächlich betrachtet - ein Experte in Sachen "Sprache". Ob er nun bewusst oder unbewusst die Regeln seiner Muttersprache beherrscht, so erscheinen sie ihm doch zu komplex und heterogen, als dass er sich von einem Computer hierzu der intellektuellen Leistung vergleichbare Ergebnisse und Funktionen vorstellen könnte. Die inzwischen kommerziell verfügbar gewordenen Ad-hoc-Lösungen auf Zeichen- und Wortebene bestätigen eher diese Vorstellungen als sie zu falsifizieren.

Auf der anderen Seite - und dies setzt sich bis in die Gegenwart fort - wurden linguistische Modelle in erster Linie - wenn überhaupt - zu Forschungszwecken in Computerprogramme umgesetzt. Im Vordergrund stand die Simulation von Sprachanalyse und -verstehen. Extreme Beispiele sind die Modelle der sog. Künstlichen Intelligenz, die bislang allenfalls auf kleine sprachliche wie physische "Welten" bezogen waren. Diese wissenschaftlich und insbesondere sprachwissenschaftlich sehr wohl begründeten Ansätze und Verfahren sind für die Lösung praktischer Probleme bislang weitestgehend ohne Wert geblieben. Um es einmal mit der Herstellung von traditionellen Wörterbüchern zu vergleichen: Wenn man ein Wörterbuch für den Schulgebrauch machen möchte, kann man sich nicht auf die Lösung der Frage der Bedeutungsdifferenzierung konzentrieren und dies an einem Beispiel (etwa dem Wort "Liebe") exemplarisch erproben. Umgekehrt reicht es z.B. für die Verwendung im Fremdsprachenunterricht nicht aus, den wesentlichen Wortschatz einer Sprache alphabetisch geordnet aufzulisten, vielmehr gehören Angaben zur Flexion, zu Wortfamilien (Komposita, Derivationen), zum syntaktischen Gebrauch, zur Betonung und Aussprache, evtl. zur Etymologie sowie Bedeutungserklärungen und Merkmale zu fachsprachlichen und stilistischen Besonderheiten dazu.

*

Wenn man - wie hier - bei der Computergestützten Texterschließung für ein stufiges Vorgehen plädiert, so muss man sich bewusst sein, dass jede Stufe (die in sich, nebenbei bemerkt, wiederum Variationen aufweisen kann, also eher als abstraktes Konzept zu sehen ist) Unvollkommenheiten und Teillösungen in sich birgt, die durch tiefere Erkenntnisse bzw. Funktionen überholt werden können. Die stufige Vorgehensweise hat jedoch den Vorteil, dass jeweils praktikable, d.h. für verschiedene Anwendungen (und letztlich immer mehr Fragestellungen) nützliche (Zwischen-)Ergebnisse erreicht werden. Im Gegensatz zu "reinen" Ad-hoc-Ansätzen (wie sie im Bereich des kommerziellen Information Retrieval häufig zu beobachten sind) sollte jedoch die stufige Vorgehensweise so weit linguistisch motiviert erfolgen, dass die erreichten Teillösungen integraler Bestandteil weitergehender Konzepte werden können. Dies unterscheidet z.B. die Vorgehensweise bei CTX von derjenigen anderer Verfahren (z.B. sei hier auf die Key-Phrase-Technik bei SEELBACH 1975 oder die Methoden des Partiellen Parsing in ROSTEK 1979 verwiesen).

*

Die systematische Entwicklung höherwertiger linguistisch motivierter Verfahren zur Computergestützten Texterschließung erfährt inzwischen zunehmend Unterstützung aufgrund zweier entscheidender Prozesse:

Es handelt sich einmal um die bekannte Miniaturisierung der Computertechnik unter Ausweitung der Speicher- und Zugriffstechniken zu größeren Datenbeständen. So ist es heute bereits praktizierbar, maschinelle Wörterbücher als Instrument der Rechtschreibhilfe und Silbentrennung auf Textverarbeitungssystemen einzusetzen (man vgl. die Entwicklungen im kommerziellen Bereich, z.B. des IBM-Schreibsystems, bei ALPHATEXT oder auch bei NIXDORF). Für diesen Bereich der Bürokommunikation werden zwar zunächst weniger komplexe Teillösungen entwickelt, mit der Ausweitung auf Textarchivierung und -retrieval in Büro und Verwaltung aber auch höherwertige (d.h. über die wortorientierte Verarbeitung hinausgehende) Verfahren einbezogen werden.

Der zweite - anhaltende - Prozess, die stetige Kostensenkung im Bereich der Computer-Hardware, verbunden mit dem wachsenden Kostenanstieg im Personalbereich, lässt (verbesserte) Verfahren einer automatischen Indexierung zunehmend attraktiver werden in den Fällen, in denen heute noch die intellektuelle Texterschließung dominiert. Allerdings ist v.a. bei höherwertigen Systemen - angesichts der Komplexität und Heterogenität natürlichsprachiger Daten - davon auszugehen, dass die maschinelle Sprachdatenverarbeitung auf lange Sicht der intellektuellen Pflege und wohl auch Kontrolle und Interaktion bedarf, wenn die Ergebnisse von Forschung und Entwicklung möglichst optimal genutzt werden sollen.

Vor diesem Hintergrund muss die im Folgenden beschriebene Entwicklung des Systems CTX betrachtet werden. Trotz eines starken maschinellen Anteils (und auch angesichts der prinzipiellen Möglichkeit, den intellektuellen Anteil bei CTX auf ein Minimum zu reduzieren) ist das System auf die computergestützte Indexierung bzw. Texterschließung ausgerichtet. Der intellektuelle Aufwand besteht im Schwerpunkt in der Lexikonpflege. Ehe jedoch auf die Funktionen des Systems im einzelnen eingegangen wird (Kap. II), soll der Rahmen, in den die Verfahrensweise einer computergestützten Texterschließung gestellt ist, allgemein eingeführt werden (Kap. I). Dies geschieht vor allem, um auch Nicht-Fachleute mit den wesentlichen Problemen eines Information Retrieval vertraut zu machen, aber auch als Vorbereitung auf die formal-inhaltliche Indexierung des CTX-Systems.

# I METHODEN UND PROBLEME DES INFORMATION RETRIEVAL

## I.1 Allgemeine Grundlagen

Die meisten maschinenunterstützten dialogorientierten IR-Systeme verwenden eine der beiden grundlegenden Techniken zum Zugriff, genauer: zur Speicherung und Verfügbarmachung von Informationen. Sie werden im folgenden als Suchbaum-Technik bzw. als Stichwortverfahren bezeichnet. Beide haben ihre speziellen Vor- und Nachteile, und die Wahl des jeweiligen Verfahrens bestimmt entscheidend die Art der Informationssuche. Dementsprechend schließen einige neuere Systeme beide Vorgehensweisen ein (z.B. "CONDOR" von SIEMENS (Banerjee (1977)) oder übernehmen einige Überlegungen der anderen Verfahrensweise innerhalb einer Technik, um daraus einige Vorteile zu ziehen.

### I.1.1 Die Suchbaum-Technik (Menü-Technik)

Die Suchbaum- oder Menü-Technik beruht auf einer Vorstrukturierung des Wissens über den gespeicherten Informationsvorrat. In gewisser Weise in Analogie zu der Ablagetechnik einer Bibliothek im klassischen Sinn ist das Wissen in einer mehr oder weniger baumähnlichen Struktur vorklassifiziert, an deren "Blättern" oder "Termini" (Begriffe) die einzelnen Informationseinheiten (im folgenden "Dokumente" genannt) zu finden sind, sozusagen wie in einem Archiv oder im Regal einer Bibliothek aufbewahrt.

Das Suchbaum-Verfahren geht zurück auf eine typische Art der Wissensdarstellung: Ausgehend von einer allgemeinen Wurzel (d.h. von einem einzigen Ausgangspunkt) werden Schritt für Schritt bis hin zu einem "(Einzel-)Begriff" Teilbereiche auswählbar, gesteuert durch eine Übersicht (Menü) über die Anordnung der folgenden Wahlmöglichkeiten (im allgemeinen heute realisiert mit Hilfe eines Bildschirmgerätes). Wenn der Benutzer auf einen Terminus oder Selek-

tionsbegriff stößt, kann er z.B. die Möglichkeit haben, in den aktuellen Text des Dokuments einzusehen, das Dokument betreffende weitere Informationen anzufordern oder Angaben zu erhalten, die mit dem Terminus selbst gekoppelt sind, wie z.B. der Zahl derartiger Dokumente, ihre Titel bzw. Verweise, die Anzahl der Seiten/Wörter in jedem Dokument, und so weiter - jeweils in Abhängigkeit von seiner Reaktion auf das vorgegebene entsprechende "Terminal-Menü".

Beispiel:

WOLLEN SIE:

(1) Weitere Informationen zum Dokument
(2) den Text des Dokuments sehen
(3) neue Begriffe eingeben
(4) den Dialog beenden

(bitte Ziffern wählen)

Die Vorteile der Menü-Technik für den Benutzer wurden besonders augenscheinlich bei der Nutzung der Suchbaumstrategie zur Einführung in ein solches System (wie z.B. das britische System PRESTEL, in der Bundesrepublik Deutschland als Bildschirmtext (BTX) bekannt). Mit einem absoluten Minimum an Stichwortsuche und praktisch ohne Vorabinformation - abgesehen von der Kenntnis der allerersten Einstiegsprozedur selbst - kann der ungeübte Benutzer sehr rasch Erfahrung sammeln mit den Mechanismen des Suchbaum-Retrieval.

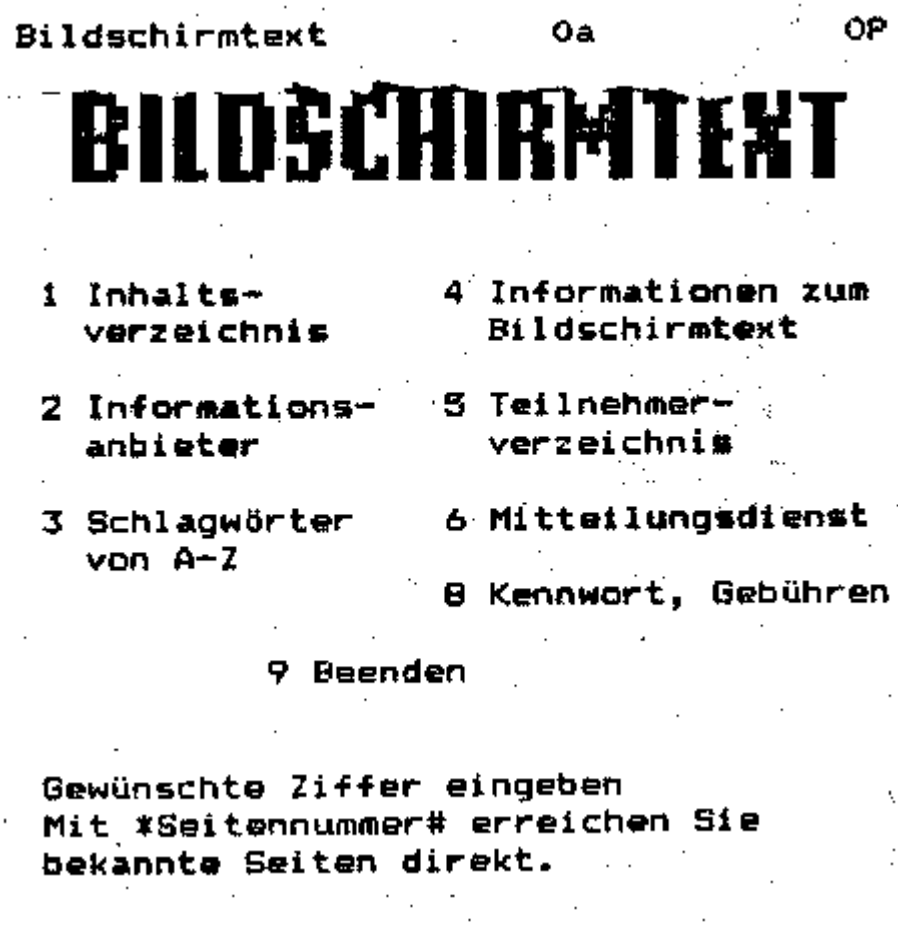


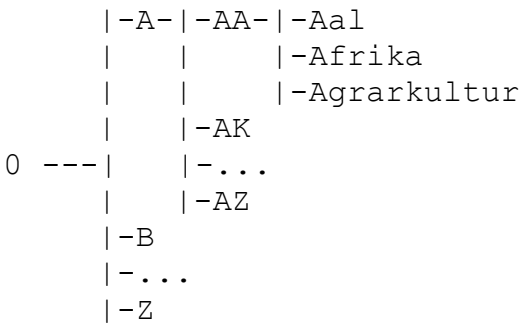

Abb. 1.1: Eingangsmenü Bildschirmtext

Aus der Sicht des Systems sind die Vorteile der Suchbaum-Technik gegenüber traditionellen Bibliotheksmethoden gleichermaßen schlagend. Ganz abgesehen von der Behebung der üblichen "Lagerhaltungs"-Probleme - dies ist ein allgemeiner Vorteil sämtlicher computerorientierter IR-Systeme - ist es relativ einfach, die Sub-Klassifikation entsprechend auszudehnen, um den pragmatischen Anforderungen jedes Teil-Archivs zu genügen - beispielsweise die Anzahl der Dokumente zu jedem Begriff unter einem vorgegebenen Maximalwert zu halten, ja sogar unter Umständen jeden Begriff auf exakt ein Dokument zu beschränken. Dem Benutzer wird eine solche Erweiterung der Sub-Klassifikation sofort durch geeignete Modifikationen der Systemvorschläge zu den ursprünglichen Begriffen deutlich gemacht.

Fortgesetzte Sub-Klassifikation ist die Hauptkonzession, die ein informationsorientiertes Suchbaum-System gegenüber dem Dokumentinhalt macht. Es ist zu betonen, daß diese Subklassifikation im Suchbaum eine Verfeinerungs-Strategie des Informationssystems im Hinblick auf den Zugang zu den Daten darstellt, nicht eine Neuanordnung der Daten selbst. Ein inhaltsorientiertes Suchbaum-System würde die Reorganisation des ganzen Informationssystems (oder eines entsprechenden Sub-Systems) vom Ausgangspunkt abwärts in Abhängigkeit von jedem neu eingeführten Dokument implizieren. Ein solches Verfahren ist theoretisch denkbar, jedoch kaum durchführbar angesichts einer überwältigenden Anzahl technischer, ökonomischer und verfahrensmäßiger Gründe.

Ein vorgegebener verfahrensmäßiger Nachteil von Menü-Systemen - gleichsam die Kehrseite der Medaille - ist die Zahl der Entscheidungsschritte (d.h. die Länge der Such-Pfade), die erforderlich ist, um einen Begriff und daraus folgend die zugehörige Information zu erreichen. Zwei typische Eigenschaften der Baumstruktur verstärken diese Tendenz. Erstens muss die Anzahl der Entscheidungen an jedem Knoten (Verzweigungspunkt) der Struktur auf einem handhabbaren Minimum gehalten werden - für gewöhnlich entsprechend der Anzahl von Entscheidungen, die ohne Schwierigkeiten (lesbar) mit einem Bildschirm-Menü dargestellt werden kann und im allgemeinen dem Benutzer erlaubt, den nächsten entsprechenden Zweig mit einem einzigen Tastendruck (bzw. in benutzerfreundlichen Systemen mit Lichtgriffel oder Fingerberührung des Bildschirmes an jeder beliebigen Stelle) anzuwählen.

Daher tendiert die Strukturbreite (d.h. die Anzahl der gleichzeitig dargestellten Auswahlwerte) zur Minimierung, und infolgedessen muss die Strukturtiefe (d.h. die Anzahl der Suchschritte) unweigerlich maximiert werden. Zweitens kann in Abhängigkeit vom Grad der "Balance" oder der strukturellen Symmetrie die Weglänge bei den Suchvorgängen beträchtlich variieren. Eine expansive Wissensstruktur neigt zum Ungleichgewicht (infolge selektiver Subklassifikation), so dass die Wege zur Information gerade in den Bereichen größter Expansion unverhältnismäßig lang werden und damit genau dort, wo eine feinere Differenzierung besonders wirksam (und nötig) ist. Als teilweise Kompensation für dieses unüberwindliche Hindernis sehen Suchbaum-Systeme im allgemeinen Abkürzungsmöglichkeiten vor, die einen direkten Zugang zu jeder gewünschten Sub-Stufe erlauben. Bei Bildschirmtext (BTX) ist dieses Abkürzungssystem einmal in einem Alphabetsuchbaum zu sehen: Ein Suchbaum "simuliert" gleichsam ein Buchstaben-Tafel-Verfahren, indem ausgehend von Anfangsbuchstaben Kombinationen über Folgebuchstaben schrittweise verfeinert wird:




Dieses Verfahren führt bei einem einzelnen Informationsbegriff (z.B. Firmenname, Schlagwort) relativ rasch zum Ziel; erst bei einem letzten Schritt muss (aus speicher-ökonomischen Gründen) sequentiell "geblättert" werden.

Ein weiteres Abkürzungsverfahren bei BTX/PRESTEL ist der Direktzugriff über ein Nummernsystem analog zu den Telefonnummern, das auch über ein systemextern verfügbares gedrucktes Angebots- bzw. Schlagwort-Verzeichnis erschlossen ist. Nachdem der Benutzer eine gewisse Vertrautheit mit der Informationsstruktur (oder Teilen davon) erlangt hat, kann er die Verfeinerung seiner Suche auch an einem individuell gewählten Sub-Menü (statt jeweils am Anfang) beginnen. In diesem Fall braucht er zusätzliche Informationen.

Für gewöhnlich liegen diese Informationen in Form eines Verzeichnisses der mnemotechnischen

Einstiegsbegriffe oder Deskriptoren mit Verweisen oder Verzweigungsmöglichkeiten vor; zusätzlich kann dieses Verzeichnis selbst als Teil der Informationsbank abgespeichert werden.

Daraus ergeben sich jedoch verfahrenstechnische Probleme, die dazu zwingen, die Struktur der Klassifikation zu erhalten:

(1) Die Vertrautheit mit der Struktur bringt nur Vorteile für den direkten bzw. abgekürzten Zugriff, wenn das System nicht in kurzen Abständen völlig neu strukturiert werden muss;

(2) Eine ständige Neu-Anpassung eines Index für direkte(re)n Zugriff in Druckform ist praktisch undurchführbar (Man denke z.B. daran, dass das Telefonbuch nur einmal im Jahr erscheint, obwohl sich bis zu ca. 30% der Einträge verändern).

Ein besonders tückischer Nachteil von Suchbaum-Systemen liegt jedoch in der Pseudo-Differenziertheit des Verfahrens. Ein spezielles Informations-"Atom" - ganz gleich, ob ein ganzer Artikel oder nur ein Abschnitt der Information - kann erfahrungsgemäß - nicht exakt und unzweideutig genau einer letzten Kategorie in einem Informationssystem zugewiesen werden. Diese Beobachtung trifft sowohl für den Makrokosmos als auch für den Mikrokosmos zu.

In früheren Jahrhunderten waren große Wissensgebiete relativ streng voneinander getrennt. In der zweiten Hälfte des 20. Jahrhunderts ergab sich eine fortschreitende Vermischung dieser Wissensgebiete, eine Überlappung der Disziplinen und ein Trend zur "ganzheitlichen" Forschung. Die Konsequenz ist, dass ein vorgegebenes relevantes Dokument (d.h. ein Dokument, das eine Antwort auf eine Benutzeranfrage darstellt) möglicherweise durch irgendeinen aus einer unendlichen Anzahl von Begriffen zugänglich ist.

Diese Begriffe können wiederum innerhalb des Informationssystems mehr oder weniger eng untereinander strukturell verknüpft (d.h. im Suchbaum weiter oder näher zueinander vermerkt) sein. Der Benutzer kann so gezwungen sein, entlang des gegenwärtigen Suchpfades weit zurückzugehen (wenn dies überhaupt gewährleistet ist) und dann einen neuen Sub-Pfad zu verfolgen - wiederum bis in große Tiefe -, um von einem potentiell relevanten Begriff (mit dem vielleicht keine relevante Information verknüpft war) zu einem anderen relevanten zu gelangen.

Auf dem mikrokosmischen Niveau ist die Verwirrung noch vollständiger. Unter welcher Kategorie soll man nach dem Begriff "Formel" suchen? Ist "Bevölkerung" ein statistischer, demographischer, mathematischer, soziologischer oder anthropologischer Begriff? Hilft es, zu wissen, dass "alkalisch" im allgemeinen als chemischer Terminus angesehen wird, wenn der Suchbereich "Textilverarbeitung" bzw. "Ackerbau" oder "Hautpflege" heißt?

Das Problem ist um so dringlicher, wenn wir bedenken, dass die Weltsicht des individuellen Systembenutzers wahrscheinlich von der des Systems (bzw. der des Systementwicklers und -pflegers) abweicht, und zwar proportional zum Grad der Spezialisierbarkeit hinsichtlich des Themas - grob gesagt zum Grad der Tiefe in der Wissensstruktur - besonders im eigenen Erfahrungsbereich des Benutzers.

Es ist ein Trugschluss, anzunehmen, dass der vermittelnde Systembenutzer (z.B. der Informationsanalytiker, der für die Einrichtung des IR-Systems für die Zwecke des Endbenutzers, z.B. die Informationserschliessung, verantwortlich ist) mehr über die Struktur von Spezialbereichen weiß

als die Spezialisten selbst wissen: aber er ist in der Regel derjenige, der Tag für Tag Entscheidungen trifft (oder treffen soll) bezüglich der Subklassifikation der Informationsstruktur. Sogar wenn man voraussetzt, dass er in der Praxis diese Entscheidungen an Bereichsspezialisten delegieren kann: werden jemals auch nur zwei Fachleute übereinstimmen hinsichtlich der Fein(sub)klassifikation ihrer (teilweise überlappenden) Interessengebiete? Oder: Wird der gleiche Spezialist bei zwei verschiedenen Gelegenheiten die gleiche Entscheidung treffen? Die Art, in der ein Themenbereich erfasst wird, wird stark beeinflusst von der Voreinstellung, um nicht zu sagen Voreingenommenheit des Spezialisten bzw. von einer speziellen Fragestellung, die einen Suchverlauf veranlasst. Eine vom Spezialisten beeinflusste Klassifikation wird schließlich einen idiosynkratischen Standpunkt widerspiegeln, wie einleuchtend dieser Standpunkt auch immer sein mag.

Wenn man diese prinzipiellen Schwierigkeiten auch nicht verkennt, so kann es - wie die Dokumentationspraxis zeigt - natürlich durchaus brauchbar-praktische Klassifikationssysteme geben. Doch selbst wo sie für die Einordnung eines Terminus vielleicht zureichend sind - sogar, dies ist vielfach zu bezweifeln -, reichen sie auch selten aus, um das Thema eines Dokuments in seinen möglichen Aspekten und Situationen präzise genug zu beschreiben.

Nun gibt es Techniken, um dieses Problem der verwischten Randzonen (und mitunter der verwischten Zentren) bis zu einem gewissen Grad zu kompensieren. Ein baumstrukturiertes Informationssystem kann z.B. viele Kopien gegebener Substrukturen enthalten. Ein besonders deutliches Beispiel ist die sog. Facettierung der Informationsstruktur. Beispielsweise ist CHEMIE (ORGANISCH) (zumindest weitgehend) identisch mit einer entgegengesetzten Substruktur des Gebietes BIOLOGIE (CHEMISCH). Zusätzlich könnten also durch Facettierung "alternative" Wissensstrukturen auf jedem Sub-Level in das System integriert werden, entsprechend den unterschiedlichen Weltsichten der Endbenutzer. Indem man Redundanz in Kauf nimmt, kann eine Wissensbasis wesentlich flexibler gestaltet werden, während gleichzeitig die einfache baumähnliche Struktur formal erhalten bleibt. Dieses Verfahren wird in der Facettenklassifikation benutzt.

Zwei Punkte sprechen dennoch gegen diese Lösung. Erstens gibt es keine theoretische Obergrenze für das Maß der Redundanz, die auf diese Weise eingebracht werden kann. Letztlich könnte nahezu jeder Begriff im System als dominierender oder dominierter Begriff in Relation zu nahezu jedem anderen Begriff fungieren - so wie es viele Möglichkeiten gibt, einen Kuchen zu zerschneiden (oder zu zerkrümeln). Des weiteren können "alternative" Wissensstrukturen ad infinitum gebildet werden und werden doch niemals den Anforderungen jedes einzelnen potentiellen Benutzers genügen. Schließlich ist unkontrollierte Redundanz kostspielig - hinsichtlich der Betriebsmittel, der Verwaltung, der Zugriffszeit und Fehleranfälligkeit. Da jede Wissensstrukturierung zu einem gewissen Grade willkürlich ist, sind Irrtümer und Inkonsistenzen nicht notwendig augenfällig.

In einer einfachen (nicht-redundanten) Struktur wird jedes Dokument einem (und nur einem) Begriff zugeordnet. In der redundanten Struktur muss ein Dokument unter Umständen gleichzeitig einer beliebigen Anzahl von Begriffen zugeordnet werden. Die Chancen für falsche Verknüpfung oder unvollständige Zuordnung der Dokumente steigen proportional zum Grad der Redundanz selbst - ebenso wie der Verwaltungsaufwand. An einem Punkt wird eine Komplexitätssättigung erreicht - je nach Anwendung kann das weit vor der Stelle geschehen, ab der die Einführung der Redundanz für irgendwelche erkennbare Flexibilität beim Retrieval sorgt - und die Informationsbasis hört einfach auf, intellektuell handhabbar zu sein.

Eine andere Teillösung kann darin bestehen, die formale Komplexität der Struktur zu steigern, d.h. vom Baum zum Netzwerk oder (noch genauer) zum gewichteten Graphen überzugehen. Eine solche Struktur kann besonders effektiv zur Vermeidung von Rückverfolgungen genutzt werden, da systemseitig Querverweise eingeführt werden können, um strukturmäßig nicht zusammengehörige Kategorien miteinander zu verbinden, z.B.

SCHMERZ von MEDIZIN (KLINISCH (THERAPIE)) mit
SCHMERZ von BIOLOGIE (HUMAN (NEUROLOGIE)).

Diese Möglichkeit kann jedoch letztlich das ganze Konzept der Informationsstruktur verändern: die Begriffe "Wurzel" und "Strukturtiefe" verlieren bei zunehmender Verwendung dieser Möglichkeit allmählich an Bedeutung, die Suche wird immer weniger vom Allgemeinen zum Speziellen gesteuert. Das Konzept der Suche nähert sich dann dem des (noch unten zu erläuternden) Stichwortverfahrens, allerdings ohne die systematischen Kontrollen, die dieses Verfahren einschließt.

In einer Graphenstruktur kann jede gegebene Sub-Struktur beliebig viele Einstiegspunkte haben, d.h. sie kann über beliebig viele Menüs erreicht werden. Diese Möglichkeit verringert in gewissem Umfang die Redundanz, die sonst in einer Baumstruktur auftritt, wo mehrfache Eingänge mittels Erzeugung identischer Kopien der Sub-Struktur simuliert werden müssen. Eine derartige Lösung nützt jedoch wenig bei teilweiser Redundanz (wenn also die Sub-Struktur nicht völlig identisch ist). Sie ist zudem kontextabhängig, d.h. das Menü, das an einem beliebigen Punkt angegeben wird, hängt ab von dem besonderen Pfadverlauf, (d.h. der "Geschichte" des Zugriffs), der selbst wieder beliebig komplex sein kann. Die Verwaltung der Rückverfolgungen wird in jedem Fall technisch noch komplizierter, besonders in einem Time-Sharing-System, also in einem System, das gleichzeitig einer Vielzahl von Benutzern zur Verfügung steht. (So lässt - wohl aus entsprechenden Gründen - Bildschirmtext nur max. drei (!) Rückschritte auf dem Menüpfad zu.) Obwohl die technischen Probleme im Grunde nicht unüberwindlich sind, ergeben sie einen beträchtlichen System-Overhead.

Von größerer Bedeutung allerdings ist, dass sich kreuzende transitive Pfade (z.B. CHEMIE (BIOLOGISCH) und BIOLOGIE (CHEMISCH)) einen nicht schleifenfreien Graphen bedingen, so dass man theoretisch nicht dafür garantieren kann, dass eine gegebene Suche jemals bei einem Begriff endet.

Es scheint, dass bei der Suchbaum-Technik die verschiedenen Lösungsansätze für die auftretenden Probleme allgemein schlimmere Auswirkungen haben als die Probleme selbst: eine einfache baum-ähnliche Struktur repräsentiert letztlich ein greifbares und stabiles Etwas, mit dem der Benutzer - sogar erfolgreich - arbeiten kann (vorausgesetzt, die Informationsmenge ist nicht zu groß).

Die ausgeprägte Inflexibilität der Suchbaum-Struktur ist v.a. günstig in Bereichen, in denen der Benutzer selbst (sowohl was die Suchstrategie bzw. -technik als auch die Inhalte angeht) nicht speziell vorgebildet ist (der sog. "casual user" bzw. der "interessierte Laie"). Er kann sogar im Laufe seiner Suche ggf. soviel mit Hilfe der Struktur einer fremden Disziplin "lernen" wie aus den gefundenen Dokumenten selbst.

So gibt es genügend wirksame Einsatzmöglichkeiten für derartige Menü-Systeme, v.a. für Sy-

steme, die den nicht speziell ausgebildeten (d.h. in seinem Themenbereich erfahrenen) Benutzer ansprechen (wir alle sind nicht speziell ausgebildet außerhalb unserer Spezialgebiete), aber auch für stark strukturierte und stabilisierte Informationsbereiche (Telefonvermittlung, Wirtschaftsregister, Einkaufskataloge), für geschlossene interne Systeme, wo der Zugang im allgemeinen zufällig ist (Bürokorrespondenz, Lagerbestandsverzeichnisse) und für die "Spitze" sehr großer Retrieval-Systeme, z.B. um dem Benutzer die Wahl zwischen getrennten Archiven oder Hauptkategorien zu ermöglichen.

Bei der intensiven Dokumentsuche jedoch, wo die Information lose oder unklar strukturiert ist, wo Zusatzfragen des Benutzers nicht so einfach zu normieren sind, wo die Dokumentdatenbank keinen festen Bestand hat und die Kenntnis des Datenbank-Inhalts nicht gegeben ist - bei diesen Bedingungen bietet die Suchbaum-'Technik wenig wirkliche Vorteile.

### I.1.2 Die Stichwort-Technik

Die zweite der beiden grundlegenden Techniken - die "Stichwort"-Technik - beruht, wie schon der Name andeutet, auf der (assoziativen) Verbindung zwischen einzelnen Dokumenten und Stichwörtern, die diese Dokumente identifizieren, repräsentieren bzw. charakterisieren. Jedes Stichwort ist demnach der Name einer Menge aller Dokumente, die mit dem Stichwort identifiziert werden. Teilmengen werden formuliert durch den Einsatz von Mengenfunktionen (mit boole-schen Operatoren) mit Argumenten, die Stichwörter repräsentieren. Zum Beispiel identifiziert

MIKROCHIP ODER CHIP

die gemeinsame Menge (Vereinigungsmenge) aller Dokumente, die entweder durch das Stichwort MIKROCHIP oder durch das Stichwort CHIP gekennzeichnet sind:

CHIP UND NICHT (KARTOFFEL ODER GOLF)

identifiziert die Menge aller Dokumente, die zwar mit CHIP, nicht aber zusätzlich mit KARTOFFEL bzw. GOLF gekennzeichnet sind (d.h., es sind wahrscheinlich Dokumente eingeschlossen, die sich auf Spielchips bzw. Computerchips beziehen).

Die Anforderungen an den Benutzer - besonders an jemanden, der nicht gewohnt ist, in komplexen Mengenausdrücken zu denken - können hoch sein. Die meisten Systeme gestatten dem Benutzer die schrittweise Eingrenzung seiner Teilmenge (d.h. sie ermöglichen den Einsatz von Mengenfunktionen auf bereits ermittelten Teilmengen). Möglichkeiten der Zeichenkettenmanipulation können vorhanden sein, mit denen der Benutzer sog. "trunkierte" Formen angeben kann (z.B.: $CHIP trifft zu auf CHIP, MIKROCHIP, COMPUTERCHIP, KARTOFFELCHIP) oder einen Bereich (z.B. MICROC$ - MICROF$ trifft zu auf MICROCOMPUTER, MICROFICHE, etc., aber nicht auf MICROJOULE, MICROMETER). Ausgabemöglichkeiten gibt es für ausgewählte Stichwörter (einschließlich den Erweiterungen obiger Ausdrücke), Teilmengenausdrücke, die Anzahl der Dokumente, die in den definierten Teilmengen enthalten sind, Informationen über einzelne Dokumente und die Dokumente selbst. Ausgefeiltere Systeme erlauben die Ausgabe von Systeminformationen, Direktzugriff (über Stichwörter) zu Dateien, Kursivschrift oder Blinken der Stichwörter im Text, Feinsuchemöglichkeiten, Updating und Korrektur der Dokumente usw. Ein sog. "Browsing", d.h. Ausweitungsmöglichkeiten gibt es für die sequenti-

elle Ausgabe von Dokumenten nach bestimmten Kriterien wie Autor, Datum der Abspeicherung und ähnliches mehr.

All das erfordert eine spezielle Interaktionssprache, die sog. Kommandosprache. Deren Grundelemente sind zwar oft schnell gelernt, eine völlige Beherrschung verlangt aber einen beachtlichen Trainingseinsatz seitens des Benutzers (und meist auch häufige und kontinuierliche Systembenutzung). Während nur wenig grundlegende Unterschiede bei der Technik verschiedener Suchbaum-Verfahren bestehen, können die Formalismen ("Sprachen") unterschiedlicher Stichwort-Systeme erheblich differieren. In jedem Fall erfordern Stichwortsysteme derzeitig zwangsläufig auch einen gewissen Grad von (Recht-)Schreibsicherheit, ein Problem, das in Suchbaumsystemen umgangen werden kann.

### I.1.2.1 Wissensorientierte Stichwortstrategien

Man kann zwei Arten von Stichwortstrategien unterscheiden. Sie differieren nicht so sehr im verwendeten Software-System als vielmehr in der Art, in der die Dokumentenbasis erfasst und strukturiert ist. Die erste Art ist - wie Suchbaumsysteme - wissensorientiert. Die Stichwörter sind dabei Ausdrücke in einer vordefinierten und vorstrukturierten Wissenswelt. Jede beliebige Anzahl solcher Ausdrücke kann (intellektuell) einer Struktur entsprechenden Begriffsinhalts zugeordnet werden bzw. einer Struktur, die nach dem (subjektiven) Urteil des Indexierers für die im Dokument enthaltene Information relevante Stichwörter aufweist. Nachdem einem Dokument die Stichwörter zugeordnet sind, werden sie in einer Datei oder Liste (Register) gesammelt, die dann wieder auf irgendeine Art mit dem gespeicherten Dokument verbunden wird. Ein praktikables IR-System wird automatisch eine invertierte Datei der Verweise aus den Stichwörtern erstellen. Beim Retrieval wird i.d.R. zunächst nur diese Datei mit den Verweisangaben zu einem Dokument verarbeitet. Der Text eines Dokuments selbst wird erst aufgerufen, wenn der Benutzer die ermittelte Teilmenge anschauen möchte.

Es ist wichtig, festzuhalten, dass in einem wissensorientierten Stichwort-System ein bestimmtes Stichwort nicht tatsächlich in dem Dokumenttext vorhanden sein muss. Umgekehrt muss ein Ausdruck, der tatsächlich im Text auftritt, nicht in der Liste der Stichwörter enthalten sein, wenn nach der Meinung des Indexierers (d.h. der Person, die nach Lektüre eines Dokuments diesem die relevanten Stichwörter zuordnet) dieser Ausdruck nicht relevant ist. Sollte zum Beispiel in diesem Text der Ausdruck INFORMATIONSSYSTEM tatsächlich nicht vorgekommen sein, könnte er nichtsdestoweniger als relevantes Stichwort vergeben werden. Andererseits können mehrmals die Begriffe CHEMIE und BIOLOGIE verwendet sein, ungeachtet der Tatsache, dass das Hauptthema nichts mit diesen Gebieten zu tun hat und diese daher vom Indexierer nicht in die Stichwortliste übertragen werden.

Ein erster (impliziter) Nachteil der wissensorientierten Strategie liegt darin, dass sie - bei aller Sorgfalt - letztlich doch eher eine willkürliche Destillierung (oder Abbildung) des Inhalts repräsentiert. Es gibt drei prinzipielle Facetten dieser Willkür:

(1) Beliebigkeit der Wissensbasis

Da die Auswahl der Stichwörter aus einer vorgegebenen Wissensbasis vorbestimmt ist, konstituiert die Basis in ihrer Gesamtheit einen Zusammenhang, durch den Dokumente aufgrund der Stichwortvergabe interpretiert werden. Wenn auch die Erwartung gerecht-

fertigt ist, dass der erfahrene (und entsprechend geschulte) Indexierer sein Urteil unter Berücksichtigung dieses Gesichtspunkts fällt, ist dies nicht in gleicher Art vom Benutzer zu erwarten (sofern dieser nicht wieder - in einer Art Vermittlerrolle - mit dem Indexierer identisch ist).

In dieser Hinsicht ist die wissensorientierte Strategie nicht besser oder schlechter als die Suchbaumtechnik, aber es gibt einige wichtige Unterschiede. Im Suchbaumverfahren kennt der Benutzer die Struktur oder lernt sie kennen; er ist zumindest durch das System "gezwungen", seine Suche nach dieser vorgegebenen Struktur auszurichten (wobei er einiges aus dem Suchpfad schließen kann). Beim Stichwortverfahren kennt jedoch der Benutzer die globale Wissensstruktur nicht. An jedem Punkt der Suche kennt er nur die Stichwörter, die er ausgewählt hat, und deren engste Umgebung (zum Thesaurus vgl. weiter unten). Diese Nicht-Übereinstimmung in der Weltsicht zwischen Benutzer und System wird beim Suchprozess kaum (oder häufig zu spät) erkannt.

(2) Willkür der Relevanz-Interpretation

Das eindeutige Ziel der Indexierung ist es, relevante Stichworte zuzuordnen und umgekehrt die Zuordnung irrelevanter Stichworte zu vermeiden. Es gibt jedoch keine einzige, klar definierte Bedeutung des Begriffs "Relevanz", keine einzige Testfunktion, die alle möglichen Benutzeranfragen und -sichten berücksichtigt. Mögliche Parameter und Auswahlkriterien sind:

| | |
|---|---|
| Themenbezogenheit: | Repräsentiert das Stichwort das (oder ein) Thema des Dokuments? |
| Häufigkeit: | Repräsentiert das Stichwort einen im Dokument häufig diskutierten Gegenstand? |
| Nützlichkeit: | Repräsentiert das Stichwort ein Thema, über das das Dokument neue oder nützliche Informationen enthält? |
| Motivation: | Repräsentiert das Stichwort die prinzipielle Motivation für die Erstellung des Dokuments? |
| Fachgebiet: | Repräsentiert das Stichwort das anerkannte Fachgebiet des Autors? |
| Existenz: | Repräsentiert das Stichwort etwas, das im Dokument erwähnt wird? |

Bei eingehender Überlegung könnten dieser Liste sicherlich weitere mögliche Auswahlkriterien hinzugefügt werden. Jede einzelne Testfunktion bzw. irgendeine Kombination aus möglichen Testfunktionen kann unter Umständen für den Zweck irgendeiner vorgegebenen Suche hilfreich sein, aber nicht notwendig besonders nützlich für andere Suchvorgänge. Eine mögliche Lösung könnte darin bestehen, das allumfassende "Relevanzkriterium" zu ersetzen durch eine Reihe von Kriterien (Themenbezogenheit, Häufigkeit, etc.) und dem Benutzer die Wahl zu lassen, welches Kriterium (ggf. als "Aspekt" einem Stichwort zugefügt) er als rele-

vant für seine Zwecke ansieht. Die Nachteile dieser Lösung liegen u.a.

- in der erhöhten Belastung des Indexierers, der nicht länger mit seinem "geübten Instinkt" arbeiten kann, sondern sich mit formalen Zwängen auseinandersetzen muss, und

- in der erhöhten Komplexität des Suchprozesses selbst.

Interessant ist im vorliegenden Zusammenhang, dass das "Existenzkriterium" das am wenigsten subjektive aller möglichen Kandidaten ist und zugleich dasjenige, das am meisten dem operationalen Kriterium der inhaltsorientierten Strategie (s) weiter unten) ähnelt. Nichtsdestoweniger gibt es einen feinen, aber fundamentalen Unterschied: "Existenz" bedeutet hier die Erscheinungsform eines Begriffs, für den das Stichwort ein Identifikationswortlaut ist - der Begriff selbst muss nicht explizit im Dokument auftauchen.

(3) Willkür im Grad der Relevanz

Selbst unter der Voraussetzung, dass eine einzelne Testfunktion formuliert werden könnte, um die Bedeutung von "Relevanz" zu charakterisieren, bleibt noch etwas, was mitunter als "Abgrenzungs-Problem" (oder als Problem der "fuzzy sets") bezeichnet wird. Ein "Thema" kann ein Hauptthema sein, ein Nebenthema oder ein eher zufälliges Thema; ein "Gegenstand" kann mehr oder weniger häufig diskutiert werden, die Information kann neu oder ziemlich neu sein, oder sie ist qualitativ und kontextbezogen von größerer oder geringerer "Bedeutung". Relevanz ist in Wirklichkeit also zusätzlich eine Frage des Maßstabs, der Skalierung.

Damit ergeben sich zwei weitere Probleme. Zum Ersten ist ein potentiell relevantes Stichwort in der Praxis entweder innerhalb oder außerhalb der Menge der Zuordnungen vorhanden, die einem so genannten Grenzkriterium ("Cutoff-Wert") entsprechen (dessen der Indexierer sich bewusst oder nicht bewusst sein kann). Aufgrund dieser Abgrenzung können geringfügige Unterschiede im Relevanzgrad genügen, um unterschiedliche Zuordnungen zu Dokumenten, die im Prinzip nahezu identisch sind, zu verursachen. Zum Zweiten muss diese Grenze bei einem spezifischen Punkt der Skala gezogen werden; ein solcher Punkt ist nicht nur selbst beliebig, die Erkennung der Grenze ist dazu noch höchst subjektiv, und daraus pflegt Inkonsistenz in der Anwendung zu resultieren.

Eine mögliche Lösung des Abgrenzungsproblems ergibt sich mit der Verbindung einer Gewichtskala mit jedem zugeordneten Stichwort, entsprechend des Relevanzgrads, der für das assoziierte Dokument zutreffend erscheint. (Der Spezialfall des Gewichts 0 wäre mit dem Fehlen des Stichworts in der Zuordnungsmenge gegeben.) Die Nachteile einer solchen Lösung sind offensichtlich: Die Skalierung bedeutet eine systematische Ausweitung der Ungenauigkeit, die sich ansonsten nur auf Grenzfälle auswirkt, und ihr Einsatz würde sowohl dem Indexierer als auch dem Endbenutzer eine beachtliche Last aufbürden. (Man kann sich heute allerdings "intelligente" maschinelle Retrievalverfahren vorstellen, die dem Benutzer ein "Browsing" in den derart erschlossenen Dokumenten ermöglichen.)

Ein zweiter Nachteil der wissensorientierten Stichworttechnik liegt im Konservatismus der Wissensgrundlage. Besonders in den Bereichen Wissenschaft und Technologie sind neue Dokumente

geeignet, Begriffe zu definieren, die dazu bestimmt sind, in der Zukunft höchst kritische Stichwörter zu werden. Gerade aufgrund ihrer Natur verändert sich eine Wissensbasis in Reaktion auf Zwänge, die unter Umständen erst nach einer Anlaufzeit in Erscheinung treten, d.h. wenn wissenschaftliche und technologische Publikationen (auch in der Terminologie) auf neue Richtungen zu antworten beginnen - Richtungen, die begründet wurden durch zukunftsweisende Publikationen, die ggf. indexiert wurden vor der allgemeinen fachlichen Einführung der neuen Schlüsselbegriffe. Die Konsequenz ist, dass entweder solche Dokumente, wiedergeholt (wie?) und ihre Verknüpfungen auf den neuesten Stand gebracht werden müssen (Re-Indexierung), oder dass sie durch genau die Stichwörter, die im Verlaufe der Terminologiebildung entstanden sind, nicht wieder aufgefunden werden können.

Eine Anzahl interessanter Punkte ergibt sich aus einem Vergleich zwischen den Möglichkeiten der Suchbaumtechnik und denen der wissensorientierten Stichworttechnik. Zunächst einmal gibt es im Stichwortsystem(ggf. abgesehen von der Einbeziehung eines Thesaurus, vgl. unten) keine nichtterminalen Begriffe: ein Retrieval mittels eines Stichworts korrespondiert mit dem Direktzugriff auf eine Dokumentmenge über ihre definierenden Charakteristika. Die Viele-zu-Viele-Abbildung zwischen Dokumenten und Stichwörtern ist also grundlegend verschieden von jeder auch nur oberflächlich ähnlichen Verknüpfung zwischen Menübegriffen und Dokumenten, und zwar aufgrund der fundamental unterschiedlichen Art der Begriffe in den beiden Typen eines Wissenssystems: Menübegriffe repräsentieren letztlich Subklassifikationen des Wissens (im Sinne einer "schrittweisen Verfeinerung"), während Stichwörter Begriffe repräsentieren, deren Assoziation mit speziellen Dokumenten bestimmt wird durch eine (weitgehend subjektive) Relevanzfunktion. Zweitens bedeutet Suchverfeinerung im Sinne des Suchbaums (Menü) den Fortschritt von Haupt- zu Unterkategorien, den Abstieg durch sukzessive Subklassifikation der Wissenswelt, wie sie durch die Menü-Struktur definiert wird, während Suchverfeinerung im Sinne der Stichworttechnik die sukzessiv präzisere Definition einer relevanten Dokumententeilmenge (z.B. durch Mengenoperationen unter Hinzunahme oder Ausschluss von Stichwörtern) darstellt: es gibt keinerlei Gerichtetheit von Klasse zu Unterklasse der Information.

Die Wissensstruktur eines Stichwortsystems kann allerdings in Form eines Thesaurus vorhanden sein (vgl. ausführlicher zum Thesaurus: I.l.2.3). In diesem Thesaurus werden Beziehungen zwischen einzelnen Begriffen bzw. Begriffspaaren definiert. Die engste Annäherung dieser Beziehung an die Gerichtetheit der Menübasis sind die Beziehungen zwischen Hypernym/Hyponym, Ganzes/Teil und Menge/Element; aber diese Beziehungen sind nicht systematisch in dem Sinne, dass ein Begriff ein grundlegendes Hypernym (oder Ganzes, oder Menge) ist, während andere dazwischenliegende oder grundlegende Hyponyme (oder Teile, oder Elemente) darstellen: die Relation ist immer binär, und die Struktur bleibt anfangslos. Weitgehend analoge Relationen zur Querverweistechnik in einem Suchbaumsystem sind bei der Stichwortvergabe Synonymie, Antonymie, Quasisynonymie, Homonymie: Der grundlegende Unterschied liegt im Gebrauch solcher Relationen (in Suchbaumsystemen: um Rückschritte zu vermeiden; in Stichwortsystemen: um andere potentiell relevante Suchbegriffe ausfindig zu machen bzw. in die Suchfrage einzuschließen) und wiederum in der Art der zwei Begriffstypen: Synonymie etc. bezeichnen begriffliche Relationen zwischen den Begriffen selbst, während der Querverweis in der Suchbaumtechnik Verbindungen aufzeigt zwischen den Wissensgebieten, die der Begriff repräsentiert.

Der Vergleich zwischen Suchbaum und Thesaurus lässt sich wie folgt zusammenfassen: Die Struktur der menügesteuerten Informationsbank ist selbst die primäre Information, die die Bank enthält, und die Begriffe haben nur die Aufgabe, die Wahlmöglichkeiten eines speziellen Such-

baums zu verdeutlichen; sie brauchen nicht einmal notwendigerweise einheitlich zu sein. Die stichwortgesteuerte Informationsbank besteht in erster Linie aus den Stichwörtern selbst, und die Struktur ist nur zufällig: Eine gegebene Thesaurusanfrage wird im allgemeinen nur mit den nächst liegenden Relationen eines Zielbegriffs verbunden, danach mit den weniger nächst liegenden, im Hinblick auf die Erweiterung der Ausbeute. Die Begriffe im Thesaurus müssen einheitlich sein, da jeder Term eine einzige Begriffseinheit aus einem vorbestimmten Kontext repräsentiert.

### 1.1.2.2 Formal-inhaltliche Stichwortstrategien

Die zweite der beiden Typen von Stichwortstrategien ist formal-inhaltlich orientiert: Stichworte, die zu einem gegebenen Dokument gehören, sind Begriffe, die formal (d.h. lexikalisch repräsentiert) innerhalb des Textes des betreffenden Dokuments erscheinen - im Gegensatz zu Termen die konzeptionell erscheinen, d.h. ggf. auch unabhängig vom Vokabular des Dokuments. Die Struktur des Begriffsinventars ist bei der formal-inhaltlichen Vorgehensweise im Normalfall nur durch die Dokumentenbasis selbst bestimmt: Die Welt ist eine Welt archivierter Dokumente, nicht eine extern erfasste Welt, zu der Dokumente relationiert werden.

WissensorientierteStichwörter haben zugegebenermaßen einen Hauptvorteil gegenüber formal-inhaltlich orientierten. Er ist auf die Nichtübereinstimmung zwischen der Menge aller relevanten (konzeptionellen) Stichwörter und der Menge aller auftretenden (formalen) Stichwörter zurückzuführen. Eine Suche auf der Basis eines vorgegebenen formalen Stichworts wird möglicherweise zu Dokumenten führen, in denen dieses Stichwort erscheint, die aber irrelevant sind für die Zwecke des Benutzers (das bedeutet "Overkill", d.h. Rückgang in der Qualität der "Precision"). Die Suche wird zur gleichen Zeit versagen beim Auffinden absolut relevanter Dokumente, wenn das spezielle Stichwort im Text explizit nicht erscheint (das bedeutet "Underkill", d.h. Rückgang in der Qualität des "Recall"). Weitestgehend kann dieser doppelte Nachteil formal-inhaltlich orientierter Stichwortsuche durch die Erhöhung der Komplexität in der Suchanfrage kompensiert werden. Im allgemeinen ist nämlich anzunehmen, dass ein Dokument, das einen bestimmten Gegenstand diskutiert, bis zu einem gewissen Punkt wenigstens einige Begriffe der Terminologie enthält, die für diesen Gegenstand relevant sind. Daher kann die Qualität des Recall verbessert (oder im Fall der technischen Literatur vielleicht maximiert) werden durch Abfragen der Elemente einer Menge, die die Vereinigung der Dokumentenmengen darstellt, die über eine Reihe geeigneter relationierter Stichwörter auffindbar sind Umgekehrt würde man erwarten, dass ein Dokument, das einige wenige Begriffe aus einem gegebenen Fachbereich enthält, weniger relevant ist als eines, das viele enthält. Daraus folgt, dass eine Suche, die den Zwischenbereich zwischen den Mengen abfragt, die über einzelne relationierte Stichwörter zugriffsfähig sind, geeignet ist, genau die Dokumente auszuschalten, in denen das Auftreten einzelner Stichwörter irreführend ist. Eine typische Endbenutzerstrategie würde entsprechend zunächst die Identifikation von Mengen (Einheiten) mit hohem Recall, aber potentiell niedriger Precision einschließen, gefolgt von der versuchsweisen Identifikation von Teilmengen dieser Zwischenbereiche mit sukzessive höherer Precision, bis das Ziel der Suche erreicht ist. Failsafe-Rückgriffe sind dabei immer erlaubt: auf die Gesamtmenge, auf Einzelbegriffteilmengen und auf Browsing-Strategien.

### I.1.2.3 Der Thesaurus

Ein Thesaurus stellt ein Inventar von Begriffen dar, die semantisch miteinander in Beziehung ge-

setzt sind. Im Zusammenhang mit jeder der beiden Stichwortstrategien ist ein effektiver Thesaurus von größter Bedeutung. Besonders in einem Gegenstandsbereich außerhalb des Spezialwissens des Benutzers ist die relevante Terminologie entsprechend weniger offensichtlich. Während der Benutzer - wie erwähnt - in einem Suchbaumsystem viel aus der Anordnung der Menüstruktur selbst lernen kann, kann der Anwender eines Stichwortsystems ohne die Hilfe eines Thesaurus lediglich aus den aufgefundenen Dokumenten selbst etwas erfahren - und gerade die besonders relevanten Dokumente enthalten genau die Information, die die Effizienz ihres Retrieval erhöhen würde: ein klassischer Fall von Henne und Ei.

Die Bedeutung des Thesaurus ist wiederum sicherlich größer für ein formal-inhaltliches als für ein wissensorientiertes Stichwortsystem, da die Wahrnehmung der Dokumentrelevanz nicht so sehr auf einzelnen, gut gewählten Stichwörtern beruht als vielmehr auf der Wechselwirkung zwischen einer Auswahl von Stichwörtern, die einen Begriffsbereich abdecken.

Wenn von einem gegebenen Stichwort angenommen wird, dass es relevant ist, wird die Einbeziehung seiner Synonyme bzw. auch der Hyponyme in die Suche die Wahrscheinlichkeit erhöhen, dass relevante Dokumente gefunden werden, ohne notwendigerweise den Einzugsbereich zu erweitern; das gleiche gilt für Teile vom Ganzen bzw. Elemente von Mengen. Synonymie und Hyponymie können für die kontrollierte Erweiterung des Einzelbereichs ausgenutzt werden, ebenso die mehr subjektiv definierten Beziehungen wie Assoziation und Quasi-Synonymie. Beziehungen wie Abkürzung/Langform und Schreibvarianz können die Qualität des Recalls eines bestimmten Stichworts steigern. Speziell definierte Beziehungen für ein ausgewähltes Teilgebiet können Charakteristika, die in einer bestimmten Textsorte auftreten; ausnutzen. Zum Beispiel wird in der diese Forschungen begleitenden Modellanwendung des Systems CTX im Bereich Datenschutz durch die Relationen IUS und REG eine Beziehung zwischen "gesetzlicher Regelung" und "Regelungsgegenstand" erzeugt (vgl. die entsprechenden Vorschläge aus DIN 1463)

AUSKUNFT (REG) AUSKUNFTSANSPRUCH
AUSKUNFTSANSPRUCH (IUS) AUSKUNFT

Nun muss der Thesaurus nicht allein auf den Vorrat an fachlich relevanten Stichwörtern beschränkt sein. Nichtfachsprachliche Ausdrücke sind im Einstiegsbereich einer Anfrage, die außerhalb des fachlichen Kompetenzbereichs eines Benutzers liegt, ggf. nützlicher als hochtechnisierte bzw. fachspezifische Begriffe; solche Ausdrücke können im Thesaurus über geeignete Relationen mit "echten", d.h. fachgebietsrelevanten, Stichwörtern verbunden eingebracht werden. Man könnte sich auf sie als "Einstiegsbegriffe" beziehen. (Nebenbei bemerkt ließen sich die "Einstiegsbegriffe" wie auch die Fachbegriffe - über spezielle Hilfen, z. B. über "Hilfsdokumente", die wie die "echten" Dokumente gespeichert sein können, erläutern bzw. definieren.) Wo der Laie nach Information über JURISTEN sucht, wird die Rechtsdokumentation vielleicht durchgehend unterscheiden zwischen NOTAR und RECHTSANWALT; JURIST selbst braucht niemals im Text des Dokuments zu erscheinen und daher auch nie als Stichwort qualifiziert zu werden. Die Terminologie, die sich auf den gleichen Begriff bezieht, variiert von Gegenstand zu Gegenstand: so ist TODESFALL als Stichwort in Lebensversicherungstexten vertreten, wohingegen Informationen über Beerdigungen von TRAUERFALL sprechen, Rechtstexte werden vielleicht konsistent auf VERSTORBENE verweisen. Ein Begriff, der in einem Zusammenhang ein Stichwort darstellt, kann in einem anderen Zusammenhang ein Einstiegsbegriff sein.

Begriffe der Umgangssprache sind im allgemeinen (v.a. bei häufigen Wörtern einer Sprache) se-

mantisch sehr ambig (hier verstanden i.S. der "lexikalischen" Bedeutungen), während sie in einer Fachsprache meist viel eingeschränkter belegt oder auch definiert sind. Zum Beispiel hat ALKOHOL als chemischer Begriff nur eine einzige Bedeutung, während in der, Umgangssprache ALKOHOL der chemische Fachbegriff, das alkoholische Getränk, das Ergebnis einer alkoholischen Fermentierung oder auch Stellvertreter für geistige Getränke allgemein sein kann. Das Problem für das IR-System besteht darin, zunächst zu entscheiden (oder vorzuentscheiden), ob der Benutzer einen speziellen oder einen allgemeinen Begriff ausgewählt hat, und in letzterem Fall, welche der vielen möglichen Bedeutungen nun gemeint ist. Das ist in Wirklichkeit jedoch lediglich ein Beispiel der allgemeinen Homonymieproblematik.

### I.1.2.4 Textanalyse und Homonymieproblematik

Im Bereich der computergestützten Inhaltserschließung von (natürlichsprachigen) Texten liegen die größten Möglichkeiten im Bereich der technischen Verbesserung formal-inhaltlicher Stichwortsysteme. Die Homonyme stellen jedoch ein Hauptproblem in jedem Textverarbeitungssystem dar. Ein BAUM in botanischer Bedeutung ist nur im übertragenen Sinn mit BAUM in mathematischer Bedeutung verknüpft, wiederum unterschieden von BAUM im genealogischen Sinn. Während der Sprachwissenschaftler homonyme Mengen extrahieren möchte, wollen dies der Botaniker, der Mathematiker oder Genealoge im allgemeinen nicht: ungelöste Homonymie beeinträchtigt natürlich die Qualität der Precision - manchmal auf die lächerlichste und (vom menschlich-intellektuellen Standpunkt gesehen) dümmste Art. Ein Dokumentenbestand, der Homonymie nicht unterscheidet, wird den Benutzer anfangs amüsieren, schließlich aber verärgern. Eine teilweise Lösung kann natürlich immer darin bestehen, Suchanfragen mit assoziierten Begriffen zu formulieren; z.B.:

BAUM (UND) WEIHNACHTEN

wird kaum Dokumente ergeben, die formale BÄUME abhandeln. Es gibt jedoch einige unvermeidbare Nachteile in der Anwendung dieser Technik, um Homonyme zu entdecken. Erstens ist sich der Benutzer im Allgemeinen nicht über die offensichtlichsten Homonyme im klaren, bis er durch Schaden klug geworden ist. Zweitens ist diese Lösung weder erschöpfend noch systematisch. Drittens ergeben sich Überspezifikationen der Suchkriterien, und viertens kann einer oder können alle der Begriffe in einem logischen Ausdruck ambig sein; z.B.:

BAUM (UND) Y,

wobei Y WURZEL, ZWEIG, BLATT, etc. bedeuten kann, wird dennoch Dokumente ergeben, die sich sowohl mit botanischen als auch formalen mathematischen BÄUMEN befassen.

Eine präzisere Lösung liegt in einem Dialogverfahren zwischen System und Benutzer: Sobald etwa ein mehrdeutiger Begriff in einer Suchanfrage auftritt, wird ein "Menü" der möglichen Bedeutungen vom System angegeben. Der Benutzer hat dabei die Möglichkeit, die gewünschte Bedeutungsvariante zu bezeichnen (Classification). Dies setzt natürlich voraus, dass bei der Indexierung entsprechend differenziert wurde. Verfeinertere Techniken für spezielle Homonymieprobleme werden im folgenden abgehandelt.

Eine ausgeprägte morphologische, syntaktische, semantische und textuelle Analyse schafft Möglichkeiten, die Aussagekraft von Stichwörtern zu erhöhen. So kann bereits die morphologische

Reduktion von Varianten eines Stichworts (Wortformen) auf eine einzige kanonische Form (z.B. Grundform) den Benutzer erheblich entlasten. Es ist v.a. bei stark flektierenden Sprachen (wie dem Deutschen) irritierend für den Benutzer, zum Beispiel für jedes Substantiv die Singular-, Plural- oder Dativ-, Genitiv-, etc.-Form in einer Suchanfrage anzugeben.

Trunkierungsmöglichkeiten, d.h. etwa Kennzeichnungen für zu vernachlässigende Zeichen nach einem Wortstamm, geben gewisse Hilfen, allerdings mit einigen Einschränkungen. BEIN$ ("$" sei das Trunkierungszeichen) führt nicht nur zu BEIN, sondern auch BEINE, BEINES, BEINEN, sondern auch zu BEINHALTEN, BEINHALTET etc. Abgesehen von diesen Fehlern versagt die (einfache) Trunkierung in allen Fällen der Wortstammveränderung durch Umlautung (HAUS : HÄUSER), unregelmäßige Konjugation (BRINGEN : BRACHTE), alternative Schreibweisen (COMPUTERSYSTEM : COMPUTER-SYSTEM) und Diskontinuitäten (EIN- UND AUSGABE, BRACHTE . . . VOR).

Viele dieser Probleme erfordern neben dem Einsatz einer ausgefeilten Technik der String-Manipulation (z.B. Trunkierung mit Abstandsangaben, Linkstrunkierung über Wortlisten) eine nichttriviale Analyse - auch wenn man dabei unterstellt, dass der Benutzer sich in jedem Fall über alle Probleme im klaren ist.

### I.1.2.5 Retrieval über den Dokumentinhalt

Retrieval des Inhalts über ein Stichwort kann theoretisch zwei Formen aufweisen: entweder können die Dokumente selbst während einer aktuellen Suchanfrage auf das Auftreten von Stichwörtern hin überprüft werden (erste Alternative) oder die Stichwörter können durch einen Indexierungsvorgang zuvor aus dem Text ermittelt werden (zweite Alternative). Die erste Alternative ist aus einigen Gründen derzeit praktisch unmöglich. Erstens wäre die Durchsicht der Gesamttexte für jeden Suchvorgang, besonders bei einem großen (und dann erst wirklich nützlichen) Dokumentenbestand, ausgesprochen ineffizient. Sogar eine ansatzweise Real-Time-Affixabtrennung käme nicht in Frage, schon gar nicht eine tiefergehende morphologische, syntaktische, semantische oder textuelle Analyse. Das Verfahren würde die Nichtanwendung automatischer Verfahren zur Thesauruspflege implizieren (da solche Möglichkeiten die vorherige Aufbereitung der Texte erforderlich machen). Dies spricht ganz klar für die zweite Alternative als (derzeit) einzig praktikable Technik, mit der Möglichkeit, die direkte on-line-Dokumentdurchsicht für Displayzwecke bzw. für bestimmte Aspekte einer Feinrecherche (auf voranalysiertem Material) zu reservieren. Dabei soll der prinzipielle Nachteil dieses Verfahrens nicht verschwiegen werden, dass nämlich neuere Erkenntnisse bzw. Verfahren der Informationserschließung auf "Altdaten", nur über eine Re-Indexierung eingebracht werden können.

### I.1.2.6 Voranalyse von Texten

Diese zweite Alternative ist verfahrenstechnisch äquivalent zur wissensorientierten Stichworttechnik: Stichwörter werden mit dem Text - entsprechend dessen Inhalt - verbunden, und diese Stichwörter repräsentieren "gemeinsam" das Dokument. Auf sie wird während der Suche aktuell zugegriffen. Der grundlegende Unterschied zwischen beiden Strategien liegt darin, dass die Relation zwischen Stichwort und Dokument im formal-inhaltlich orientierten System eher durch eine (formale) Existenzfunktion als durch eine Relevanzfunktion bestimmt wird: das bedeutet, die Funktion ist eher formal als subjektiv definiert und daher - dies ist von besonderer Bedeutung für die Konzeption des JUDO-Projekts und des daraus resultierenden CTX-Verfahrens - algorith-

misch bestimmbar.

Daraus erhebt sich die Frage, ob der intellektuelle Arbeitsprozess der Ermittlung inhaltlich relevanter Stichwörter (wissensorientierte wie inhaltliche Stichwortstrategie) auf der Basis einer formal-inhaltlich orientierten Stichwortstrategie effektiv ersetzt und dabei durch automatische oder halbautomatische Hilfsverfahren ökonomisch verwertbar werden kann. Diese Frage beschäftigte das JUDO-Projekt. Der im Rahmen des Projekts angebotene Lösungsweg wird noch eingehender besprochen werden; zunächst sei ein genereller Vergleich der Stärken und Schwächen manuell-intellektueller und automatischer Techniken im allgemeinen vorangestellt, auch um die Methode der Computerunterstützung zu begründen.

### I.1.2.7 Formal-inhaltliche Stichwortermittlung (Deskribierung)

Manuell-intellektuelle Methoden weisen hier nur wenige Vorteile gegenüber automatischen auf. Wenn man sich klarmacht, dass nahezu jedes Wort, mit Ausnahme der Präpositionen, Konjunktionen und Pronomina (und evtl. der gebräuchlichen Adverbien) ein Stichwort ist (oder sein kann), würde sogar der einfache Ausweg der automatischen Aussonderung von Nichtstichwörtern über eine relativ kurze Stoppwortliste (vielleicht mit 50 Wortformen wie DER, IN, UND) immerhin über 50% des Ballastes unterdrücken. Was verbleibt kann des weiteren bei der intellektuellen Durchsicht der computergenerierten Stichwörter ausgemerzt werden, aber das ist nicht einmal unbedingt notwendig.

Zufällige Stichwörter (z.B., gesetzt den Fall, die Konjunktion WOHINGEGEN fehlt in der Stoppwortliste und erscheint in einem Dokument) können einfach als Überlauf vorhanden sein, ohne für den Benutzer einen systematischen Nachteil darzustellen. (Dies ist die "übliche" Praxis in großen formal-inhaltlich orientierten Informationsbanken.) Ein derartiges Pseudo-Stichwort würde zunächst automatisch in die Verweisliste eingeordnet, dort beim nächsten routinemäßigen "Hausputz" gefunden, automatisch aus jedem einzelnen Dokumentrecordfile, in dem es als Stichwort gespeichert wurde, gelöscht und gleichzeitig (ebenfalls automatisch) in die Stoppwortliste eingefügt.

Es gibt statistische Methoden - primär etwa auf Wortlänge und Worthäufigkeit (z.T. auch auf Abfragefrequenzen) basierend - die eine ähnliche Aufgabe erfüllen, ohne eine intellektuell gepflegte Stoppwortliste zu benutzen, obwohl diese Methoden das Risiko der nicht auszuschließenden Aussonderung auch wichtiger Stichwörter bergen. Die Ermittlung von Stichwörtern mittels syntaktischer Klassen (d.h. z.B. Beschränkung auf Substantive, Verben und Adjektive) könnte einen höheren Prozentsatz an Ballast aussondern, aber es ist zu fragen, ob die verbesserte Treffer-Quote den System-Overhead rechtfertigt, den eine Analyse bis zum gewünschten Perfektionsgrad einschließen würde.

### I.1.2.8 Pflege der Stichwortliste

Die Stichwortliste eines formal-inhaltlich orientierten Systems ist unausweichlich weniger stabil als die eines wissensorientierten Systems: Die Aufnahme jedes neuen Dokuments bringt eine gewisse Anzahl von Stichwörtern, die für zukünftige Suchläufe in die vorhandene Liste integriert werden müssen: die Pflege der Stichwortliste (insbesondere bei zusätzlicher Begriffsrelationierung in einem Thesaurus) wird so zum integrierten Bestandteil des Prozesses der Datenbankerweiterung, und zwar mehr als ein einmaliger Vorgang während des Systemaufbaus. (Ähnliches

gilt ja auch für die computergestützte Übersetzung, wo eine Wörterbuch- und Terminologiepflege als selbstverständlich angesehen werden muss.)

Mit der Listen- oder Thesauruspflege sind zwei völlig voneinander unabhängige Probleme verbunden: einmal die Kontrolle und Handhabung neuer Information, zum anderen die Bestimmung der Relationen zwischen den Begriffen eines expandierenden Thesaurus. Die Steuerung des Informationsflusses innerhalb eines geschlossenen Systems ist die eigentliche Stärke des Computers, und intellektuelle Verfahrensweisen können sich nicht im mindesten mit der Schnelligkeit und Genauigkeit des Computers messen. Der Zeitpunkt der Eingabe eines neuen Begriffs ist per definitionem sein erstes Auftreten in einem gespeicherten Dokument. Wie immer die Vorgehensweise an diesem Punkt aussieht - ob ein Begriff durch die Maschine erfasst wird, durch den Menschen oder durch eine Interaktion Mensch-Maschine - wenn er erst einmal im System ist, kann seine Einführung in die Stichwortliste bzw. den Thesaurus oder in jeden anderen Datenbereich (z.B. das Stammformenlexikon) vollkommen automatisch erfolgen.

Die Bestimmung von Thesaurusrelationen ist im Gegensatz dazu in erster Linie ein intellektueller Vorgang, aber der Computer kann hier ebenfalls entscheidende Hilfen bieten. Verschiedene Relationen lassen sich relativ einfach aus dem Begriffsinhalt selbst ermitteln:

(1) Syntaktische Relationen

| Typ | Beispiel |
|---|---|
| ADJ + SUB | AKADEMISCHER GRAD |
| SUB + Genitivattribut | WEITERGABE DER/VON DATEN |
| SUB + PRPattribut | RECHT AUF/ZUR LÖSCHUNG |
| SUB-Anreihung | ANSTALTEN UND/ODER STIFTUNGEN |

(2) Morphologische Relationen

| Typ | Beispiel |
|---|---|
| VRB - SUB | HEIZEN - HEIZUNG |
| ADJ - SUB | SCHOEN - SCHOENHEIT |
| ADJ - VRB | KRAEFTIG - KRAEFTIGEN |

Dazu kommen Relationen, die sich selbst wieder aus bereits vorhandenen Relationen ableiten lassen:

(1) Kommutative Relationen
(2) Assoziative Relationen
(3) Inverse Relationen
(4) Transitive Relationen.

Diese formalen Relationen können beispielsweise automatisch aus den vorhandenen Relationen generiert werden (im Fall der assoziativen und der transitiven Relationen ist jedoch eine zusätzliche intellektuelle Überprüfung erforderlich).

### I.1.2.9 Reduktion von Wortformen auf Grundformen

Die Rückführung von Flexionsformen (Wortformen) auf eine sie alle repräsentierende Form (z.B. Grundform) stellt noch eine verhältnismäßig einfache Forderung dar. Verschiedene morphologische Erscheinungen sind hier (z.B. im Deutschen) zu berücksichtigen: Affixabtrennung (FORT-BESTEHEN:BESTEHT ... FORT), Stammveränderung bei Starken Verben (BRINGEN: BRACHTE), Umlaut bei der Pluralbildung von Substantiven (HAUS:HÄUSER), vollständige flexionsbedingte Graphemfolgenverschiedenheit (GUT:BESSER). Für die computergestützte Reduktion solcher morphologischer Erscheinungen auf ihre Grundform sind drei Verfahrensstufen zu absolvieren:'

(1) Affixabtrennung,
(2) Reduktion der Wortstammveränderung,
(3) Lemmatisierung.

Diese stufige Vorgehensweise hat einen guten Grund: Lemmatisierung setzt die Reduktion auf den ursprünglichen Wortstamm voraus, diese wiederum für gewöhnlich die Identifikation eines Affixes. Ein gegebenes System sollte also ein-, zwei- oder dreistufige Wortformenreduktionen (zu denen Zwischenstufen und -varianten denkbar sind) durchführen.

zu (1)
Unter Affixabtrennung wird in diesem Zusammenhang die (wiederholte) Anwendung von Reduktionsregeln auf die Zeichenketten, die die Wortlauteinheiten bilden, verstanden. Der "Rest", also die Zeichenkette, die nach der Anwendung aller möglichen Abtrennungsregeln verbleibt, wird als die "Grund"-Wortform festgehalten. Es erfolgt kein Wörterbuchzugriff, alle Entscheidungen müssen allein aufgrund von Zeichenmustern getroffen werden (vgl. zu diesem Themenkomplex KUHLEN 1977a, ZIMMERMANN 1972e).
Manche Grundformen können nicht mit Hilfe von Regeln aus flektierten Varianten rekonstruiert werden. Dann bestehen zwei Möglichkeiten: entweder alle möglichen Reduktionsformen zu generieren und dabei einen äußerst redundanten Overhead in Kauf zu nehmen oder eine willkürliche Auswahl zu treffen, in der Gewissheit, dass einige Stichwortformen unrichtig und evtl. unkorrigierbar zugeordnet werden.

Die Reduktion der Wortstammveränderung erfolgt durch Abgleich der Ergebnisse des Affixerkennungsverfahrens mit einem Stammformenlexikon. Zeichenketten, die nicht im Lexikon erscheinen, können als falsch zurückgewiesen oder als evtl. fehlende Wörterbucheinträge gekennzeichnet werden. Der Wörterbuchzugriff produziert allerdings einen entsprechenden System-Aufwand.

zu (2)
Volle Lemmatisierung setzt einen gewissen Grad an syntaktischer und v.a. semantischer Analyse voraus. Während das Ergebnis der vorhergehend beschriebenen Phase die vollständige Menge möglicher Interpretationen jeder aufgetretenen Wortform darstellt, trifft dieses Verfahren die konkrete Auswahl aus den möglichen Lesarten (Bedeutungen) einzelner Wörter (d.h. Zeicheninhalten an der Sprachoberfläche), abhängig von syntaktischen und kontextuellen Informationen. Derartige Analyseverfahren werden weiter unten exemplarisch behandelt, wobei zugleich Voraussetzungen und Grenzen im Zusammenhang der Entwicklung der Verfahrensbausteine von CTX dargestellt werden.

Orientiert an der vorangehenden Darstellung sollen zunächst automatische und manuelle Methoden linguistischer Analyse miteinander verglichen werden. Es muss dabei vorausgeschickt werden, dass die Maschine etwas nicht kann und vermutlich niemals können wird: Sie wird keine ernsthafte Konkurrenz für eine intensive Text-Analyse durch einen Menschen darstellen, schon allein in Bezug auf die erreichbare Qualität der Ergebnisse: letzten Endes erfordert Analyse Verstehen, und selbst die fortschrittlichsten Techniken der Künstlichen Intelligenz können nicht annähernd die Erfahrung und Verstehenstechnik sogar eines relativ ungeübten menschlichen Gehirns erreichen.

Gegen diesen heute unausweichlich erscheinenden Schluss aber muss eine Anzahl praktischer Überlegungen in die Waagschale geworfen werden

(1) Die von Menschen durchgeführte Text-Analyse ist langsam und arbeitsaufwendig, selbst bei äußerster Ausnutzung von Maschinenunterstützung (geteilter Bildschirm, Textscanning mittels Cursor, Online-Updating usf.). Ein IR-System von einigermaßen praktisch nutzbringendem Umfang und Laufzeitverhalten kann der Mannschaft eines ökonomisch besetzten Textdatenbank-Pflegeteams harte Konkurrenz machen.

(2) Die Schwäche der maschinellen Analyse liegt nicht so sehr in der Lösungsfindung als in der Auflösung von Ambiguitäten: der umgekehrte Fall trifft auf die menschliche Analysefähigkeit zu. Ein maschinelles System kann so ausgelegt werden, dass ein Computerirrtum sich im Überangebot (und infolgedessen im Rückgang der Precision) äußert, ein menschlicher Fehler bedeutet i.d.R. ein Unterangebot (also einen Rückgang des Recall). Wenn man zu wählen hat, ist es vermutlich besser, irrelevante Dokumente zu finden, als relevante nicht zu finden.

(3) Es ist unmöglich, zu erwarten, dass der analysierende Mensch sich formal aller linguistischen Strategien bewusst ist, die er für die Entwicklung seiner (nichtsdestoweniger sehr genauen) Lösungen einsetzt, oder gar, dass er diese Strategien explizit angeben kann. (Dies kann z.B. bei der Ausbildung von Indexierern zum Problem werden.) Im Gegensatz dazu kann die maschinelle Analyse, die erforderlich ist, um eine relativ zuverlässige und eindeutige morphologische Analyse zu garantieren, auch dazu verwendet werden, weiterreichende Ergebnisse (z.B. die ermittelten Strukturen) in die weiteren Verfahren einzubringen.

(4) Eine systematische, konsistente menschliche Analyse erfordert auf lange Sicht einen (ständigen) beträchtlichen Schulungsaufwand.

### I.1.3 Verwendung maschineller Indexierungsverfahren bei formal-inhaltlich organisierten IR-Systemen

Der gegenwärtige und absehbare Stand der automatischen Analyse natürlichsprachiger Texte, besonders im Bereich Information und Dokumentation, scheint für eine Analyse mit systematisch maschinenunterstützten Verfahren, ergänzt um intellektuelle Nachbereitung oder Interaktion, zu sprechen: Die Frage ist, welcher Perfektionsgrad der maschinellen Analyse in einem IR-System erreichbar ist.

Die Abtrennung von Affixen ist sicherlich eine Minimalanforderung. Eine Wörterbuchintegration wird a priori gerechtfertigt durch die Forderung, dass ein Wörterbuch in Form eines Thesaurus in

jedem Fall in das IR-System eingebracht werden sollte. Die Integration höher qualifizierter (z.B. syntaktischer, semantischer und textueller) Analyseverfahren zu erproben erscheint nützlich. Der Aufwand für höher qualifizierte Verfahren ist relativ hoch und wächst progressiv. Ob eine solchermaßen verfeinerte Analyse zu rechtfertigen ist, hängt letztlich von fünf Kriterien ab:

1. dem Standard-Umfang der aktuellen Dokumente,
2. der Dichte des Retrieval-Verkehrs,
3. der texttypologischen Stabilität des Dokumentenbestands, -
4. den kritischen Grenzen des Qualitätsretrievals,
5. dem Nutzen der hinzugekommenen strukturellen Information.

### I.1.3.1 Dokumentumfang

Falls die Dokumente (bzw. der bei dem Retrieval eingrenzbare Dokumentenbereich) standardmäßig sehr umfangreich sind (ganze Artikel, Bücher oder größere Kapitel, etc.), erscheinen formal-inhaltlich orientierte Stichwortverfahren im Prinzip weniger geeignet als wissensorientierte Verfahren. Ohne ein Selektions- oder Gewichtungsverfahren würde allein schon die Anzahl der inhaltlichen Stichwörter, die mit jedem einzelnen Dokument verknüpft werden müssten, unverhältnismäßig groß. Die sich ergebende Tendenz eines Überangebots müsste zudem kompensiert werden durch gleichzeitige Erhöhung der Komplexität der logischen Ausdrücke, die bei der Suche anzuwenden sind. Die Techniken, die hier in Verbindung mit formal-inhaltlich orientierter Stichwortermittlung behandelt wurden, könnten in diesem Fall als Hilfsmittel für die intellektuelle Zuordnung (d.h. ggf. eine Selektion i.S. einer Kondensierung) von konzeptionellen Stichwörtern genutzt werden. (Zu einer weiteren Möglichkeit, dies zu kompensieren, vgl. die Konzeption der Mehrwortausdrücke und Komplexen Deskriptoren in der Modellentwicklung CTX). Wo die der Indexierung zugrundeliegenden Dokumente (Texte) jedoch normalerweise aus einzelnen Paragraphen oder wenigen Sätzen bestehen (z.B. bei Abstracts, Titeln, Kapitelüberschriften), erscheint die größere Bereichsabdeckung durch das formal-inhaltliche Stichwortverfahren gerechtfertigt. Da die intellektuelle Ermittlung formaler Stichwörter in jedem Fall mit grösserem Aufwand verbunden ist, werden sich dabei computergestützte Methoden im Vergleich zu intellektuellen Verfahrensweisen letztlich als kostengünstiger erweisen.

### I.1.3.2 Verkehrsdichte

Unter "Dichte" wird hier grob die Anzahl der innerhalb einer gegebenen Zeitspanne gefundenen relevanten Dokumente verstanden, geteilt durch die Anzahl der Dokumente in der Datenbank. Je höher diese Retrievaldichte ist, desto niedriger sind die Kosten pro gefundenem Dokument und daher um so größer der Betrag, der wirtschaftlich betrachtet für die Pflege der Datenbasis als Ganzes aufgewendet werden kann. Dies ist ein augenfälliges, doch auch kaum abschätzbares Kriterium, da es schwer zu trennen ist von anderen Faktoren wie z.B. Retrievalqualität (im allgemeinen: befriedigende Retrievalergebnisse) und dem Grad, in dem dieser "Satisfaktionskoeffizient" durch die Einführung hoch entwickelter linguistischer Verfahren verbessert wurde.

### I.1.3.3 Stabilität

Stammformenlexika, semantische Wörterbücher und Thesauri sind bei weitgehend thematisch begrenzten Datenbanken und Analysestrategien weniger unbeständig als bei unbegrenzten, fachgebietsunspezifischen Datensammlungen. Der Grad der nachfolgenden Pflege in Reaktion

auf jedes neue Dokument nimmt im entsprechenden Verhältnis zur Erweiterung der Datenbank selbst ab , und der Grad, indem der vorhandene lexikalische Bestand sich als ausreichend erweist, steigt im gleichen Maß. Das Ergebnis ist: rasch sinkender Aufwand bei der Systemwartung, weniger Nachkorrektur, weniger menschliche Eingriffe pro Dokument, je enger der Themenbereich einer Datenbasis begrenzt ist.

Unter Begrenzung kann die Beschränkung auf einen gegebenen Fachbereich (z.B. Datenschutz, Lebensmittelchemie, Hochfrequenztechnik, Energiegewinnung, Textilverarbeitung, Bergbau) oder auf einen bestimmten formalen Typ von Dokumenten (z.B. Abstracts, Rezensionen, Zeitungsartikel, Arbeitspapiere etc.) oder beides verstanden werden.

### I.1.3.4 Kritische Grenzen

Die Qualität des Retrieval mag die Kosten in bestimmten kritischen Anwendungsbereichen, besonders dort, wo die menschliche Gesundheit oder Sicherheit betroffen ist (z.B. medizinische Diagnose, Verteidigungssysteme, Kriminalstatistiken, Notdienste) wohl aufwiegen. Die einzige Frage ist dann, ob mechanische Verfahren absolut bessere Ergebnisse erreichen können als menschliche oder traditionelle Methoden, und sie überwiegt alle anderen Überlegungen im Hinblick auf Manpower, Training und Kosten. Aber gerade hier - wo also intellektuelle Feinanalyse und Datenaufbereitung letztlich qualitativ höherwertige Ergebnisse liefern kann - stellt sich die Frage, oh nicht durch ergänzende Anwendung maschineller Verfahren die Qualität des Retrieval nicht zusätzlich noch gesteigert werden kann.

### I.1.3.5 Anwendbarkeit

Es wurde darauf hingewiesen, dass die strukturelle linguistische Analyse nicht zum Selbstzweck werden sollte. Dies gilt auch für eine zu starke Verfeinerung der Lemmatisierung (Disambiguierung). Im Einzelfall mögen derartige Verfahren hilfreich sein, etwa bei der Feinrecherche, d.h. während einer letzten Abprüfung der bereits ausgewählten Dokumententeilmengen auf die potentiell relevanteren einzelnen Dokumente. (Der Weg, auf dem dies möglich ist, wird eingehender im Zusammenhang mit der Entwicklung der Verfahrensbausteine von CTX erläutert werden.)

Weitere Möglichkeiten außerhalb der automatischen Indexierung und des Dokumentretrieval selbst gibt es für die weitere Auswertung der Analyseergebnisse: automatische Übersetzung (MD), Frage-Antwort-Systeme, computergestützte Inhaltsanalyse, automatische Zusammenfassung, stilistische Analyse, Autorzuordnung, Forschungen in Computerlinguistik und Künstlicher Intelligenz. Eine zweckgeteilte Dokumentdatenbank verteilt die Kosten ihrer Instandhaltung unter eine größere Gruppe von Anwendern.

All das setzt natürlich voraus, dass ein grundlegendes, existentielles Kriterium schon erfüllt ist: dass die Verfahren, die in diesem ersten Abschnitt behandelt wurden, nicht nur experimentell realisiert sind, sondern auch im Rahmen eines IR-Systems eingesetzt werden. Das, um es ins Gedächtnis zu rufen, ist eines der wesentlichen Ziele des Projekts JUDO. Der Grad, bis zu dem im Forschungsprojekt dieses Kriterium erfüllt wurde, und die Form, in der die Durchführung vonstatten ging, werden im folgenden Abschnitt behandelt.

## I.2 CTX - ein Modellsystem für ein formal-inhaltliches Stichwortverfahren

Auf der Grundlage der Darstellung in Kapitel I.1 kann das Modellsystem CTX in groben Zügen folgendermaßen beschrieben werden:

- Ziel ist es, formal-inhaltlich orientierte Stichwörter zu Dokumenten aus den Texten zu erschließen, die für das Retrieval gespeichert werden;
- ergänzend ist ein Thesaurussystem eingebracht, um die bekannte Schwäche (textbezogener) formal-inhaltlicher Verfahren wenigstens partiell zu kompensieren.

CTX integriert ein automatisches linguistisches Analysesystem, dessen Zweck es ist, die Dokumente zu erschließen durch

- eine vollständige Lemmatisierung der Textwortformen (Grundformenermittlung) und
- die Bereitstellung geeigneter morphologischer, syntaktischer und semantischer Informationen zur Präzisierung der Deskribierung durch präkoordinierte Begriffe.

Das gleiche Verfahren der linguistischen Analyse kann heim Retrieval zur "Kompatibilisierung" eines Suchproblems auf eine natürlichsprachige Problembeschreibung angewendet werden. Das Modellsystem benutzt dabei die Komponente zur automatischen Sprachanalyse des Saarbrücker Übersetzungssystems (SUSY) (MAAS 1978). CTX verfügt über mehrere Schnittstellen zu IR-Systemen. Realisiert wurden Umsetzungen in TELDOK (TELDOK (1978)) und GOLEM (Siemens (1981)). Diese Testanwendung in unterschiedlichen IRSystemen hat sich unter zwei Gesichtspunkten als nützlich erwiesen: Einmal konnte die Übertragbarkeit von CTX auf verschiedene Systeme gezeigt werden, zum anderen wurde eine Basis für die Beantwortung der Frage geschaffen, inwieweit eine von einem (kommerziellen) IR-System unabhängige Software der Effizienz der CTX-Methodologie Beschränkungen auferlegt. Diese Punkte werden später zusammen mit einer Beschreibung der IR-System-Unabhängigkeit des Modellsystems ausführlich diskutiert.

### I.2.1 Die linguistische Analyse

Die Einbeziehung eines bestehenden Verfahrens zur Sprachdatenverarbeitung in CTX ist in der Praxis problematischer als in der Theorie. Theoretisch gesehen ist ein Analyse-Modul so "gut" wie das andere, vorausgesetzt, dass die Qualität der beiden vergleichbar ist: CTX selbst ist v.a. methodisch relativ unabhängig von der Art, mit der ein Verfahren diese linguistische Analyse ausführt, und von den linguistischen und computertechnischen (analysestrategischen) Theorien, die dieses Verfahren beinhaltet.

Der Output aller modernen MÜ-Analyse-Verfahren ist im wesentlichen eine Baumstruktur (in welcher Darstellung auch immer) einer gegebenen "Übersetzungseinheit" (obwohl sie im allgemeinen nicht notwendig die Länge eines Satzes haben muss). Die ermittelten strukturellen Abhängigkeiten - entweder syntaktisch oder semantisch - müssen allerdings für die Zwecke von CTX ausreichen, um Stichwörter und andere, komplexere Ausdrücke (Verbindungen von Stichwörtern) zu identifizieren (vgl. unten: Komplexe Deskriptoren). CTX liefert für diesen Teil nur zwei Schnittstellen: eine erste vor Beginn des Analyseverfahrens zur Prä-Formatierung des Textes in die Form, die für das Analyseverfahren benötigt wird, eine zweite am Ausgabe-

Ende dieses Verfahrens, um die Analyseergebnisse in das Format zu überführen, das für die weitere Verarbeitung durch eigenständig entwickelte Bausteine "CTX-intern" gefordert wird. Da das Modellsystem mit der spezifischen Eingliederung von Verfahrensbausteinen des SUSY-Systems geplant worden ist, treten diese Schnittstellen nicht explizit in Erscheinung oder genauer: der Programmteil zur Textvorbereitung des SUSY-Systems ist vollständig bezüglich der Inputkonvention von CTX als Schnittstelle übernommen worden; zusätzlich ist versuchsweise die (modell-)interne Struktur von CTX in Übereinstimmung mit den Analyse-Konventionen der SUSY-Analyse entwickelt worden:

Beide Schnittstellen sind in diesem Fall weitgehend "Null-Schnittstellen". Man kann sich jedoch eine Produktionsumgebung denken, in der unterschiedliche Analyseverfahren verwendet werden - z.B. für verschiedene Quellsprachen - und in der nicht-leere Schnittstellen relativ oberflächiger Art viele der willkürlichen Unterschiede zwischen den Systemen ausscheiden. Man kann sich weiterhin den zusätzlichen Anschluss von einem oder mehreren hoch entwickelten Textverarbeitungssystem(en) vorstellen, abhängig von der Verfügbarkeit oder der Wahl des Benutzers; der Output müsste dabei für den Input in das Analyseverfahren lediglich umformatiert werden. Schließlich könnte eine technische Umformatierungs-Schnittstelle zum Ausgleich von Unterschieden in der Hardware gefordert werden zwischen der Ausrüstung, die für die Textvorbereitung verwendet wird und der, die zur folgenden Verarbeitung benutzt wird.

In der Praxis entstehen Probleme jedoch in erster Linie aus den Bedingungen des Entwicklungsfortschritts, die mit der linguistischen Analyse zusammenhängen. Erstens variieren die Anwendbarkeit und die Qualität der Software zur linguistischen Analyse je nach der Sprache. Während die (strukturellen) Entsprechungen für die wesentlichen europäischen Sprachen (Deutsch, Französisch, Russisch, Englisch) relativ groß sind, sind sie weniger groß bei den nicht-europäischen Sprachen; und obwohl vorgesehen ist, dass wenigstens alle offiziellen EG-Sprachen in absehbarer Zukunft adäquat vertreten sind, kann man das gleiche nicht für das Afghanische, Bantu, Magyar usw. behaupten. Zudem versprechen Entwicklungen wie EUROTRA (das geplante System der Europäischen Gemeinschaft zur Maschinellen Übersetzung (MÜ) eine gewisse Regelmäßigkeit in der Übersetzungsqualität für häufig verwendete Sprachen, wogegen die Qualität der Übersetzung und die Anwendbarkeit von Analysewörterbüchern und Wortdatenbanken außerhalb des dominierenden Bereichs westeuropäischer Sprachen diesen Zielsetzungen nicht in diesem Maße gerecht werden können.

Das Problem der Verwendbarkeit von Analyseverfahren kann ein Problem für irgendeine Anwendung von CTX bedeuten oder nicht; es ist abhängig von der Sprache oder den Sprachen, in denen die gespeicherten Dokumente geschrieben sind: ein einsprachiger Leser eines Artikels wird die Schwierigkeit z.B. nicht spüren. Schwieriger und problematischer ist es, dass der Anstoß für die stetige Ausdehnung und Verfeinerung linguistischer Analyse-Strategien primär aus dem Bereich der MÜ kommt, und dass sich daher die Software-Entwicklungen auf diesem Gebiet nicht auf die Analyse als solche konzentriert haben, sondern eher auf den Typ und den Grad der Analyse, der notwendig ist, um eine Übersetzung zu leisten. Um zu beurteilen, wie diese Orientierung die Verwendung von Komponenten der MÜ-Analyse bei CTX grundlegend beeinflusst, soll zunächst der Aufbau von MÜ-Systemen näher betrachtet werden.

### I.2.2 Exkurs: Maschinelle Übersetzung (MÜ)

Man kann MÜ-Systeme (für den Zweck dieser Untersuchung) in vier Hauptkategorien untertei-

len, die sich durch ihren Modus der äußeren System-Organisation unterscheiden. An dieser Stelle konzentriert sich die Betrachtung auf die Wirkung, die dieser Aufbau in jedem einzelnen Fall auf die Analyse-Strategie hat. In jedem Fall muss man zwischen theoretischen Möglichkeiten und historischer Entwicklung unterscheiden: Parallel zu der Entwicklung der globalen Organisation von MÜ-Systemen (im groben in der Reihenfolge, in der sie unten vorgestellt werden) gibt es gleichermaßen signifikante Entwicklungen in den theoretischen Modellen und in der Computer-Methodologie, die mit individuellen MÜ-Projekten verbunden sind. Dies hat den Effekt, dass viele der früheren Systeme ihr eigenes volles Potential linguistischer Analyse nicht realisiert haben, wie man es heute rückblickend verstehen kann. Es ist sehr gut möglich, dass in der Zukunft ähnliche Feststellungen bezüglich der Systeme, die heute in Entwicklung und Erprobung sind, getroffen werden.

Die erste Kategorie umfasst Systeme, in denen eine Übersetzung direkt von der Quellsprache in die Zielsprache vorgenommen wird, in der Art einer "Alles-in-einem"-Übersetzungskomponente. Der Input in diese Komponente kann in gewissem Maße prä-editiert sein, im typischen Fall, um eine morphologische Reduktion der Oberflächenstruktur der Quellsprache zu erreichen; ähnlich kann der Output der Überführung eine Synthese der Oberflächen-Wortform in der Zielsprache erfordern. Für den Zweck dieser Betrachtungen jedoch kann man die prä-editierten und post-editierbaren Zustände als Quellsprache und Zielsprache ansehen, bzw.:

Quellsprache => Überführung (Transfer) => Zielsprache

Die Übersetzung besteht im Großen und Ganzen aus einer kontrastiven Ad-hoc-Wiederzusammenstellung der Oberflächenelemente, wobei kein systematischer Versuch linguistischer Analyse unternommen wird.

Bei einem solchen Modell brachten die großen Hoffnungen der 50er und 60er Jahre - die so genannte "Erste Generation von MÜ-Systemen" - ausgesprochen unbefriedigende Ergebnisse, was zu der Erstellung des ALPAC-Berichts führte (ALPAC 1966); dessen Kritik an der MÜ hatte die fast vollständige Aufgabe der Forschung und Entwicklung für mehr als ein Jahrzehnt zur Folge, insbesondere in den USA.

Für die Betrachtungsweise des JUDO-Projekts ist ein solches System in jedem Fall von geringem Interesse. Es gibt keine erkennbare Analyse-Komponente bzw. keine systematische Analyse-Struktur. Was in einem solchen System an Analyse existiert, wird in Form von Übersetzungsäquivalenten ausgedrückt.

Die zweite Kategorie ist aus der ersten entstanden; die Übersetzungskomponente ist in zwei Unterkomponenten aufgespalten, die als "Analyse" und "Synthese" bezeichnet werden, mit einer zwischengeschalteten Vermittlungsstruktur. Diese Vermittlungsstruktur - in der Literatur oft irreführend als "Zwischensprache" bezeichnet - stellt das Resultat der (gewöhnlich) wesentlichen Analyse der eingegebenen Übersetzungseinheit dar:

Quellsprache => Vermittlungsstruktur => Zielsprache
Analyse Synthese

Folgende Beobachtungen sind hier von Bedeutung: Erstens ist theoretisch die Entwicklung eines Systems auf dieser Basis möglich, das die modernsten und höchstentwickelten syntaktischen, se-

mantischen und textuellen Analysetechniken verwendet; die historische Wirklichkeit zeigt jedoch, dass solche Systeme modernisierte Versionen von Systemen der ersten Generation waren und auch weiterhin bleiben und deshalb primär syntaxbezogen sind, bestenfalls mit der Verwendung semantischer Analyse bei günstigen Gelegenheiten, wenn die syntaktische Analyse mehrdeutige Ergebnisse erbringt. Zweitens wird die Analyse mit der Kenntnis von und in Übereinstimmung mit den Erfordernissen einer bekannten Zielsprache vorgenommen: Sie ist nicht so sehr das Ergebnis einer quellsprachlichen Analyse als vielmehr ein Mittelpunkt im Übersetzungsprozess. Bei jeder gegebenen Quellsprache wird die Vermittlungsstruktur auch bei jeder Zielsprache völlig verschieden sein.

Natürlich ist eine Analysekomponente, die auf solchen Prinzipien aufbaut, alles andere als optimal für die Zielsetzungen des JUDO-Projekts, dessen Erfordernisse v.a. in einer befriedigenden Analyse der Quellsprache liegen.

Da die Analysestrategie der Quellsprache weitgehend von zielsprachlichen Überlegungen bestimmt wird, sind Systeme der Kategorien l und II im allgemeinen als "bilinguale" Übersetzungssysteme bekannt. Im Gegensatz dazu sind Systeme der Kategorien III und IV "multilingual".

Zur dritten Kategorie gehören die Übersetzungssysteme, die eine "wirkliche" Zwischensprache oder eine maschinenlesbare Kern-Darstellung verwenden, deren Form und Inhalt (weitgehend) unabhängig von den Besonderheiten einer der einbringbaren Quellsprachen sind (Kanonische Repräsentationssprache). Für alle n Sprachen, die durch ein solches System abgedeckt werden, kann man den Übersetzungsvorgang in folgendem Diagramm darstellen:

```
SPRACHE 2  =>                          => SPRACHE n-1
                ZWISCHENSPRACHE
SPRACHE 1  =>                          => SPRACHE n
```

Die Pfeile von der Sprache zur Zwischensprache bedeuten Analysemoduln, und die von der Zwischensprache zur Sprache stehen für Synthesemoduln. Jede Übersetzung von zwei Sprachen ist daher in einem solchen System ein zweistufiger Vorgang, oberflächlich gesehen identisch mit dem eines Systems der Kategorie II; der Unterschied wird deutlich, wenn man die zwei Diagramme für die Übersetzung von einer Sprache in viele vergleicht:

```
                QUELLSPRACHE
                     =>
                INTERLINGUA
     =>              =>              =>
ZIELSPRACHE 1  ZIELSPRACHE 2  ZIELSPRACHE n
```

Der Vorteil dieser Art von MÜ-Analyse für die CTX-Konzeption ist, dass die Zwischensprache notgedrungen eine ziemlich vollständige Analyse erfordert - wenn man von einer Zwischensprache ausgeht, in der alle linguistischen Konstruktionen explizit markiert sind und nicht nur diejenigen, die für die Übersetzung zwischen einem gegebenen und bekannten Paar von Sprachen notwendig sind. Für die Zwecke der automatischen Indexierung wäre eine Zwischensprache ideal. Immerhin ist es in MÜ-Kreisen eine unbestrittene Tatsache, dass die Analyse"tiefe", die für eine gute Übersetzung unbeschränkter natürlicher Sprache nötig ist, den Status eines sprachverstehen-

den Systems nahe-, wenn nicht sogar gleichkäme. Der gegenwärtige und absehbare Zustand der maschinellen Analyse natürlicher Sprache(n) ist von dieser Forderung allerdings noch weit entfernt. Daher ist hochwertige interlinguale Übersetzung in der MÜ nur für Systeme mit begrenztem Input praktikabel, Systeme, in denen der quellsprachliche Text syntaktisch und semantisch stark eingeschränkt ist, um linguistische Phänomene zu umgehen, die eine automatische Analyse nicht bewältigen kann.

Daraus folgt, dass das Analyse-Modul eines solchen MÜ-Systems nur in einem CTX-System von Nutzen wäre, in dem die gleichen Input-Beschränkungen auf die gespeicherten Dokumente angewendet werden könnten. Obwohl sinnvolle Anwendungen auf dieser Basis vorstellbar sind, ist es das Hauptanliegen des Forschungsprojekts JUDO, einen Anschluss an Informations-Datenbanken zu schaffen, somit auf Daten Bezug zu nehmen, die keine von außen auferlegten Beschränkungen besitzen, d.h. das Verfahren ist auf beliebige Texte (Dokumente) einer Sprache ausgerichtet.

Die vierte Kategorie umfasst alle Systeme der sog. "zweiten und dritten Generation" d.h. alle unbeschränkten multilingualen Systeme "nach ALPAC" ab den späten 60er Jahren. Hier ist die Übersetzung ein dreistufiger Prozess, mit zwei vermittelnden Strukturen, die durch eine Transferkomponente miteinander verbunden sind.

Der Zweck der Einführung eines Transfer-Moduls ist die Bereitstellung eines Mediums für sprachpaarspezifische Analyse, die getrennt ist von der zentralen, von der Zielsprache unabhängigen Analysekomponente. Für die Zwecke der MÜ bewirkt dies eine Reduktion der Redundanz - der Großteil der Analysekomponente muss für jede Quellsprache nur einmal implementiert werden - und zudem eine Reduktion der Analysetiefe, die für die Erreichung einer zufrieden stellenden (brauchbar-praktischen) Übersetzung notwendig ist, auf ein realistisch erreichbares Maximum. Während die (sprachpaarbezogene) Transfer-Komponente bei passender Gelegenheit und ad-hoc spezifische kontrastive Merkmale eines einzelnen Sprachpaares ausnutzen kann, kann die Analysekomponente im wesentlichen von zielsprachlichen Überlegungen unabhängig sein.

Betrachtet man jetzt die Analysekomponenten von MÜ-Systemen isoliert, kann man die jeweiligen Vorzüge der vier Kategorien in Verbindung mit ihrer Verwendungsmöglichkeit für ein CTX-System gegeneinander abgrenzen:

Kategorie I:
Sowohl wegen der theoretischen Unzulänglichkeiten als auch wegen der historischen Begrenztheit solcher Systeme ist die Anwendung für die Texterschließung wenig geeignet und sinnvoll. Das CTX-System benötigt zumindest eine maschinenlesbare Struktur, die explizite Einzelheiten grammatischer Formen enthält.

Kategorie II:
Trotz der prinzipiell starken Beschränkungen kann die Kategorie der Analyse-Strategie für eine Anwendung in dem vorgestellten CTX-System nicht von der Hand gewiesen werden. In der Praxis benötigt man eine gewisse minimale Kernanalyse, um wenigstens eine grundlegende Repräsentation der Struktur von Analyseergebnissen zu erreichen und um eine systematische Synthese zu ermöglichen: Diese Kernanalyse ist primär quellsprachenorientiert. Zweitens findet man ein hohes Maß an Ähnlichkeit zwischen Sprachen - nicht nur formale Ähnlichkeit, sondern auch Ähnlichkeit im Sinne von sprachlichen Universalien: eine linguistische Erscheinung, die innerhalb eines Sprachpaares eine Übersetzungsschwierigkeit darstellt, wird zumeist für jedes Sprach-

paar ein Problem bedeuten. Drittens beseitigt bereits die Trennung der Analyse von der Synthese viel von der Unterschiedlichkeit in der Behandlung eines gegebenen Phänomens in der Quellsprache. So muss z.B. die (indirekte) syntaktische Beziehung zwischen MILCH und FÜR BABIES in "diese Milch wurde speziell für Babies zubereitet" entdeckt (und gespeichert) werden, was immer auch das Äquivalent in der Zielsprache sein mag. Wie stark die Form der Speicherung dieser Information in der Vermittlungsstruktur auch von zielsprachlichen Besonderheiten beeinflusst werden mag, so bleibt der Inhalt dieser Struktur doch ausreichend, um im folgenden die Stichwörter BABY und MILCH einander zuzuweisen.

Kategorie III:
Von allen Kategorien wird diese von den Beschränkungen der Realität gegenüber der Theorie am weitesten negativ beeinflusst. Ein Analyseverfahren dieser Art wäre für ein CTX- System ideal: wenn eines zur Verfügung stünde. In der Praxis existiert jedoch kein solches (anwendungsnah entwickeltes) Verfahren, das für eine adäquate Übersetzungsqualität in einem völlig unbeschränkten MÜ-System genügen würde. Die Analyseverfahren von MÜ-Systemen mit begrenztem Input sind, für ihren Teil, künstlich eingeschränkt: Ohne beträchtliche Erweiterungen sowohl der Lexika als auch der inhärenten Analyse-Kapazität (was vermutlich aufgrund formaler mathematischer Einschränkungen nicht möglich ist, die für die Besonderheiten der restringierten Syntax der verarbeiteten Sprachen entworfen sind und somit für die Komponenten der MÜ-Analyse selbst) sind sie für die computergestützte Indexierung in absehbarer Zukunft kaum anwendbar.

Kategorie IV:
Analyseverfahren dieser Kategorie haben Vor- und Nachteile. Ihr Vorteil gegenüber solchen der dritten Kategorie ist, dass sie verfügbar sind, und zwar für immer mehr Sprachen: Diese Kategorie stellt das gängige Wachstumsgebiet vollständiger MÜ. Ihr Vorteil gegenüber den Verfahren der zweiten Kategorie ist, dass die Zwischendarstellung der Analyseergebnisse standardisiert und unabhängig von zielsprachlichen Überlegungen ist. Gerade diese Standardisierung birgt eine gewisse Gefahr (vom Standpunkt der MÜ-Strategie), da ein vorkommendes Phänomen bis zu einer gewissen Tiefe analysiert werden muss, wenn dessen abschliessender Transfer in irgendeine potentielle Zielsprache nicht trivial ist. Dies ist für eine Anwendung in der computergestützten Indexierung jedoch ein großer Vorteil. Andererseits setzt das Vorhandensein einer Transferkomponente voraus, dass Teile der Analyse in gewissem Maße nachgestellt sind, d.h. dass manche linguistischen Erscheinungen zum ersten Mal im Kontext eines spezifischen Sprachpaares analysiert (oder vollständig analysiert) werden und nicht vorher. Für Zwecke der MÜ würde die Auflösung von Mehrdeutigkeiten nur in einem Deutsch-X-Verfahren für eine Sprache X versucht werden, in der diese Mehrdeutigkeit für eine korrekte Übersetzung aufgelöst werden muss - anderenfalls nicht, wie es für die meisten, wenn nicht sogar alle europäischen Sprachen der Fall ist. Für die (vertiefte) computergestützte Indexierung jedoch ist die Bedeutungs-Disambiguierung notwendig, z.B. um die Relationierung eines Begriffes sicherzustellen.

Aufgrund der letzten vorgestellten Überlegungen könnte es so aussehen, als seien MÜ-Analyseverfahren für die computergestützte Indexierung uninteressant, insbesondere aufgrund eines gewissen Interessenwiderspruchs zwischen den Zielen der Indexierung und der MÜ. Diese Inkompatibilität sollte jedoch nicht zu hoch bewertet werden: Die Analyseprobleme sind bei beiden Anwendungen überwiegend gleich, und das Hauptanliegen der computergestützten Indexierung liegt in der Erweiterung und nicht im Ersatz der Ergebnisse des Analyseverfahrens. Diese Erweiterung baut auf IR-orientierten, "systeminternen" Informationen auf, die in jedem Fall einem MÜ-System nicht prinzipiell zur Verfügung stehen, und stellt ein Analyse-Verfahren dar, das für

die Indexierung selbst dann ausgeführt werden müsste, wenn es ein theoretisch ideales linguistisches Analyseverfahren gäbe. Allgemein betrachtet stellen die realisierten Verfahrensbausteine von CTX in weiten Teilen, etwa bei der Bildung komplexer Deskriptoren, selbst eine Sprachsynthese (allerdings in einer Art "Kunstsprache") dar. Ähnliches gilt auch für den Transferteil: Dort, wo die (standardmäßig) gelieferten Ergebnisse für die Indexierung nicht ausreichten, wurde in Teilbereichen eine Art "Nachanalyse" (wenn man so will - ein Transferprogramm) entwickelt. Diese Erweiterungen werden im nächsten Abschnitt im Anschluss an eine Beschreibung der spezifischen linguistischen Analysekomponente, die in die Verfahrensbausteine von CTX integriert ist, besprochen.

### I.2.3 Das Saarbrücker Übersetzungssystem (SUSY)

Mit Unterstützung der Deutschen Forschungsgemeinschaft wurde an der Universität des Saarlandes ein Sonderforschungsbereich zur Elektronischen Sprachforschung (SFB 100) eingerichtet. Im Rahmen des Teilprojekts A2 des SFB wird ein multilinguales MÜ-System entwickelt. Dieses System - bekannt als SUSY (Saarbrücker Übersetzungssystem) - ist im wesentlichen einzuordnen unter der vierten Kategorie der MÜ. Es beinhaltet Analyse-, Transfer- und Synthese-Komponenten. Die Analyse-Komponente, auch bekannt als "Saarbrücker Automatische Textanalyse", stellt ein wesentliches Teilverfahren von CTX dar. Sie besteht aus einer Reihe von Moduln, die technisch als "Operatoren" (Programmbausteine) realisiert sind. Im Einzelnen sind dies (SALEM 1980):

(1) eine Textaufbereitung zur Satzanalyse: Nach Eingabe des zu analysierenden Textes wird dieser in "Sätze" zerlegt, die Einzelwörter werden separiert u.ä.m.

(2) eine Morphologische Analyse (Wörterbuchsuche): Zu den eingegebenen Wortformen werden alle morphologisch möglichen "Lesarten" erzeugt auf der Basis der Abtrennung von Endungen und der Suche im Stammwörterbuch (ggf. mit Kompositum-Zerlegung). Somit sind - bezogen auf Einzelwörter - auch alle potentiellen Grundformen ermittelt.

(3) eine Disambiguierung syntaktischer Mehrdeutigkeiten: Es handelt sich dabei um eine vorwiegend wahrscheinlichkeitsorientierte Gewichtung individueller syntaktischer Lesarten (Wortklassenfolgen) auf der Basis distributioneller (morphosyntaktischer) Beschränkungen und quasi-statistischer Information (d.h. unter Zugrundelegung der Vorkommens- und Übergangswahrscheinlichkeit jeder potentiellen Wortklassenpaar-Kombination, auf der Grundlage empirischer Textanalysen und analysestrategischen Erfahrungswerten).

(4) eine Segmentierung in syntaktische Subsätze (Haupt- und Nebensätze): Ein Satz wird zerlegt in (wiederum potentielle) Satzsegmente (z.B. Haupt-/Nebensatzgrenzen) auf der Grundlage der Interpunktion und des morphosyntaktischen Kontextes der jeweiligen Lesarten.

(5) eine Nominalgruppen-Analyse: Nominalgruppen werden ermittelt (Strukturaufbau); die möglichen Beziehungen zwischen Nominalgruppen (Koordination, Attribution) werden aufgeführt.

(6) eine Verbalgruppen-Analyse: Verbalgruppen werden ermittelt, die verbalen Kerne für die Satzgliedermittlung werden bestimmt. Ggf. werden hierbei 'ungrammatische' Lesarten getilgt.

(7) eine (syntaktische) Komplement-Analyse: Nominalgruppen werden Verbalgruppen zugeordnet und außerdem werden Haupt- und Nebensätze koordiniert. Auch hier ist die Tilgung 'ungrammatischer' Lesarten möglich.

(8) eine Semantische Disambiguierung: Anhand eines "semantischen Lexikons" werden Bedeutungsmehrdeutigkeiten erkannt und mittels geeigneter Regeln zu vereindeutigen versucht. Grundlage ist vorwiegend die Übereinstimmung kasusähnlicher Merkmale bei Nominalgruppen in Abhängigkeit vom Valenzrahmen des Verbs. Zugleich werden mehrwortige Begriffe (feste Wendungen) unter Benutzung des semantischen Regel-Lexikons identifiziert.

Die syntaktischen Verfahrensschritte (v.a. 3-7) identifizieren die linguistische Struktur und schalten fortschreitend Mehrdeutigkeiten aus, die zunächst aufgrund der morphologischen Einzelwort-Analyse als möglich angesehen wurden. Für die Zwecke der Indexierung mit CTX reicht im Prinzip die Analyse bis einschließlich der Nominalgruppen aus, um die so genannten "Komplexen Deskriptoren" zu ermitteln. Im Gegensatz zu den beschriebenen prinzipiellen Anforderungen sind die implementierten Verfahrensbausteine CTX dabei mit einer Art "Sicherheitsnetz" versehen, so dass auch Analyseergebnisse verwendet werden können, bei denen nicht alle im Prinzip nötigen Verfahrensschritte erfolgreich durchlaufen werden. Dies kann im Einzelfall zu unaufgelösten Mehrdeutigkeiten führen oder aber - da aus praktischen Gründen (v.a. zur Reduktion von Computer-Rechenzeit) nicht alle "Lesarten" verwertet werden - zu einem gewissen (natürlich möglichst gering zu haltenden) Teil auch zu Fehldeutungen.

Die semantische Analyse greift auf ein separates semantisches Wörterbuch zu, dessen Zugriffselemente weitgehend mit einer Grundform des morphosyntaktischen Wörterbuchs identisch sind. Anhand des semantischen Regel-Lexikons werden zunächst Mehrdeutigkeiten in der Bedeutung eines Wortes (innerhalb einer Wortklasse) identifiziert. Die dazu aufgeführten Regeln sollen - soweit im syntaktischen Rahmen und aufgrund der vorgegebenen semantischen Merkmale möglich - die Disambiguierung derartiger Mehrdeutigkeiten bewirken. In diesem Zusammenhang werden auch die sog. "Festen Wendungen" ermittelt.

Die semantische Disambiguierung wird - v.a. aufgrund der oft zu allgemeinen Satzinformation - die allein als Kontext einbezogen werden kann - von der vorliegenden Teilkomponente des Saarbrücker Übersetzungssystems (SUSY) nicht erschöpfend ausgeführt (die "Auflösungsquote" liegt bei derartigen Mehrdeutigkeiten derzeit noch unter 50%). Im Rahmen des Projekts JUDO wurde daher zusätzlich eine eigene Strategie entwickelt, um weitere Disambiguierungen durchzuführen. Hierzu gehört v.a. die Einbringung und Ausnutzung einer Angabe zur Vorkommenswahrscheinlichkeit einer Bedeutungsvariante in einem (jeweils zu definierenden) Fach- oder Themenbereich. Im Test befindet sich auch ein Verfahren, bei dem Begriffsrelationen (zwischen bedeutungsdifferenzierten Wörtern) herangezogen und zur Disambiguierung auch über die Satzgrenzen hinaus verwertet werden können.

### I.2.4 Erstellung von Deskriptoren

Die Ergebnisse der Satz- und Textanalyse bilden den Input für das Verfahren zur Extraktion von formal-inhaltlichen Stichwörtern aus dem Text (Dokument).

Da - der Zielsetzung nach - diese Stichwörter als Repräsentanten für den Inhalt eines Textes/Dokuments gelten und zudem bei einer Mehrdeutigkeit der Zeichenkette an der Sprachoberfläche

geeignete Maßnahmen zur Bedeutungsdifferenzierung vorgenommen werden, lassen sich diese Stichwörter auch als Deskriptoren verstehen - wenn man sich auch über den Unterschied im Vergabeverfahren gegenüber der intellektuellen Indexierung bewusst sein muss.
Die Extraktion der Deskriptoren erfolgt unter Heranziehung eines weiteren zu diesem Zweck erstellten und gepflegten Lexikons.

Dieses 3. Lexikon-Teilsystem enthält zusätzliche Informationen fachgebietsspezifischer Art. Auf diese Weise ist es z.B. prinzipiell möglich,

- die Vergabe von sinntragenden Deskriptoren auf fachgebietsrelevante zu beschränken;
- mehrwortige Deskriptoren zu identifizieren;
- durch die sprachanalytischen Verfahren nicht differenzierte semantische Mehrdeutigkeiten über fachgebietsspezifische Angaben zur Wahrscheinlichkeit des Auftretens einer Bedeutungsvariante aufzulösen.

Deskriptoren, die mithilfe dieses Lexikons gewonnen werden, werden als "Einfache Deskriptoren" bezeichnet. Sie umfassen also einfache und zusammengesetzte Wörter ebenso wie Feste Wendungen.

Mehrdeutige Wörter werden zusätzlich mit allen sprachlich möglichen Bedeutungsvarianten verzeichnet. Jede Variante ist mit einer Definition, einer Paraphrase oder mit beidem erklärt. Dabei erhält jede Bedeutungsvariante - wie erwähnt - eine Gewichtung nach der Wahrscheinlichkeit ihres Vorkommens in diesem Fachgebiet.

### 1.2.5 Präkoordination (Komplexe Deskriptoren)

Das Ergebnis der automatischen Sprachanalyse ist unter mehreren Gesichtspunkten nützlich:

- Alle Textwortformen werden auf zugehörige Grundformen zurückgeführt (kanonisiert): Auf diese Weise sind alle möglichen Formvarianten eines Deskriptors mit einem einzigen Lexikoneintrag abgedeckt.
- Die (syntaktische) Distributionsklasse (Wortklasse) jedes Deskriptors im Text ist bestimmt worden, so dass wortklassenübergreifende (syntaktische) Homographie aufgelöst wird.
- Die explizit gekennzeichnete syntaktische Struktur erlaubt die Rekonstruktion von Komposita und mehrwortigen Deskriptoren, auch wenn Elemente im Text diskontinuierlich sind.
- Die syntaktische Information bietet die Möglichkeit, "Komplexe Deskriptoren" zu erkennen und zu bestimmen.

Diese letzte im Rahmen des Forschungsprojekts entwickelte Variante der Deskriptorerstellung beruht auf folgenden Überlegungen:

Untersucht man die Struktur mehrwortiger Begriffe, z.B. in einem Register, so lassen sich v. a. drei Typen erkennen:

(a) die Struktur "Adjektiv und Substantiv", z.B.

JURISTISCHE PERSON
PERSONENBEZOGENE DATEN

(b) die Struktur "Substantiv und attributives Substantiv", z.B.
ANSTALT DES OEFFENTLICHEN RECHTS

(c) die Struktur "Substantiv und präpositionales Attribut", z.B.
BUNDESBEAUFTRAGTER FUER DATENSCHUTZ

Es ist nun nicht immer leicht - auch für einen Experten in einem Fachgebiet - zwischen "lexikalisierten" Mehrwortbegriffen und übrigen, eher thematisch-zufälligen bzw. üblichen begrifflichen Verknüpfungen zu differenzieren. Aus diesem Grunde wurde eine allgemeine Möglichkeit der textbezogenen Präzisierung bzw. Präkoordination von Einfachen Deskriptoren bei Vorliegen bestimmter, gängiger syntaktischer Relationen geschaffen, die als Bildung sog. "Komplexer Deskriptoren" bezeichnet wurde.

Diese Art der oberflächigen "Präkoordination" ist in erster Linie ausgerichtet an den heute üblichen, am Markt angebotenen Systemen zum Information-Retrieval (z.B. GOLEM, DIRS-GRIPS, UNIDAS und STAIRS). Es wäre ebenso denkbar, derartige (syntaktische) Strukturinformationen als solche aufzubewahren und erst zum Retrievalzeitpunkt zu nutzen. In gewisser Weise stellt die Feinrecherche bei GOLEM eine solche Möglichkeit dar; in dem bei SIEMENS zum Ende der 70-er Jahre in Entwicklung befindlichen System CONDOR ist in der Tat eine ähnliche Strategie zum Retrievalzeitpunkt implementiert worden, die sog. "Relevanzfunktionen" (vgl. WIELAND 1979).

Um die Möglichkeit der Präzisierung im Rahmen des Forschungsprojekts demonstrieren zu können, wurde jedoch der in dem gegebenen Zusammenhang leichter praktizierbare Weg der Erweiterung der Deskriptorenvergabe zum Indexierungszeitpunkt gewählt.

Ein Komplexer Deskriptor besteht dabei aus einem Paar von Elementen, deren jedes - wie erwähnt - ein Einfacher Deskriptor ist. Diese Elemente stehen in einer der folgenden oberflächensyntaktischen Beziehungen zueinander:

| | | |
|---|---|---|
| (1) | Adjektiv-Attribut plus Substantiv | (ADJ-Relation) |
| (2) | Substantiv plus Genitiv-Attribut | (GEN-Relation) |
| (3) | Substantiv plus Präpositional-Attribut | (PRP-Relation) |
| (4) | Substantiv plus beigeordnetes Substantiv | (KON-Relation) |
| (5) | Verb plus Akkusativkomplement | (AKK-VRB-Relation) |
| (6) | Verb plus zugehöriges Modalverb | (MOD-VRB-Relation) |

Ein Komplexer Deskriptor wird in der Deskriptoren-Kette als ein Buchstaben-String vermerkt, der die (beiden) Einfachen Deskriptoren enthält, die ggf. durch einen Relator in der Form eines einzelnen Buchstabens (P, G oder K, je nach Art der Relation) voneinander getrennt sind. In den folgenden Abschnitten werden die Bedingungen für diese Relationen beschrieben.

ADJ-Relation

Eine ADJ-Relation wird automatisch generiert, wenn ein Adjektiv als Attribut zu einem Substan-

tiv identifiziert wird. So werden bei dem Satz

"Die deutsche Delegation erreichte eine einstimmige Entscheidung" die Komplexen Deskriptoren

DEUTSCHE DELEGATION
EINSTIMMIGE ENTSCHEIDUNG

vergeben (neben den (Einfachen) Deskriptoren DELEGATION, DEUTSCH, ERREICHEN, EINSTIMMIG und ENTSCHEIDUNG). Die während der Satzanalyse ermittelte syntaktische Struktur ermöglicht ggf. die Erkennung auch diskontinuierlicher Elemente. Die Phrase "private und vertrauliche Dokumente" liefert u.a. die Deskriptoren PRIVATES DOKUMENT und VERTRAULICHES DOKUMENT. Ähnlich liefert "öffentliche Einrichtungen und Stellen" die Deskriptoren OEFFENTLICHE EINRICHTUNG; OEFFENTLICHE STELLE u.s.w.

Derartige ADJ-Relationen lassen sich grundsätzlich auch aus anderen syntaktischen Oberflächenstrukturen generieren. So bietet sich an, auch prädikative und im Relativsatz nachgestellte Adjektive in ADJ-Relationen zu transformieren, wie z.B. in

"die Entscheidung war einstimmig"
- EINSTIMMIGE ENTSCHEIDUNG

"ein Dokument, das privat und vertraulich ist"
- PRIVATES DOKUMENT
- VERTRAULICHES DOKUMENT.

Die Ergebnisse des zugrunde gelegten Analyseverfahrens reichen prinzipiell dazu aus, diese Erweiterung zu ermöglichen. Dies konnte jedoch im Projektzeitraum nicht mehr eingebracht werden.

<u>GEN-Relation</u>

Grundlage für die Erzeugung eines Komplexen Deskriptors ist die Relation zwischen einem Substantiv und einem Genitiv-Attribut; das Attribut im Genitiv ist dabei nachgestellt, der Name der Relation wird dabei zur Relatorkennung (G) reduziert. So liefert

| | "Unternehmen der Presse" |
|---|---|
| den Komplexen Deskriptor | UNTERNEHMEN G PRESSE |

Zusätzlich bewirkt die Attribuierung eines Nomens mithilfe der Präposition "von" die Zuweisung der GEN-Relation. So liefert

| | "Wiedergewinnung von Daten" |
|---|---|
| den Komplexen Deskriptor | WIEDERGEWINNUNG G DATUM |

Es ist festzuhalten, dass die GEN-Relation - wie alle Relationen Komplexer Deskriptoren - rein syntaktischer Natur ist: es wird Z.B. nicht versucht, "echte" Possessiva von unechten zu unterscheiden, wie in

"Unternehmen der Presse" "Freiheit der Presse"

"Wiedergewinnung von Daten" "Qualität von Daten"
"Mitglied der Mannschaft" "Arbeitsbelastung der Mannschaft"

Wie bei den ADJ-Relationen kann die GEN-Relation auch zwischen diskontinuierlichen Konstituenten erkannt werden. So liefert z.B.

"Speicherung und Wiedergewinnung von Daten"

die beiden Komplexen Deskriptoren

SPEICHERUNG G DATUM
und WIEDERGEWINNUNG G DATUM

"Wiedergewinnung von Medizinischen und anderen hochvertraulichen Daten" liefert

WIEDERGEWINNUNG G DATUM
MEDIZINISCHES DATUM
ANDERES DATUM
HOCHVERTRAULICHES DATUM

PRP-Relation

In Analogie zur Erzeugung eines Komplexen Deskriptors vom Typ GEN-Relation wird das Nomen eines Präpositionalattributs zu seinem übergeordneten Nomen zugeordnet. Der "Name" der Präposition wird dabei reduziert zur Relatorkennung (P). So liefert die Phrase

"Zugang zu Informationen"
den Komplexen Deskriptor ZUGANG P INFORMATION

Es kann gelegentlich von Nachteil sein, dass individuelle Präpositionen bei der Relationierung nicht unterschieden, sondern alle auf den Relatornamen "P" reduziert werden. In manchen (wenn auch seltenen) Konstruktionen kann die Präposition sehr wohl von Bedeutung sein, um eine Suchfrage semantisch einzugrenzen. Dagegen wurde davon ausgegangen, dass Präpositionen in Präpositionalattributen in den europäischen Sprachen z.T. sehr bedeutungsarm sind - in den romanischen Sprachen noch weit mehr als in den germanischen - so dass die Präpositionen sowohl stark redundant als auch oft frei austauschbar sind. Es erschien daher wenig sinnvoll, vom Benutzer die explizite Anführung jeder einzelnen Präposition zu erwarten, um ein Dokument zu finden, das einen Komplexen Deskriptor mit einer PRP-Relation enthält.

Eine attributive Nominalphrase, die durch die Präposition "von" eingeleitet ist, bewirkt, wie oben erwähnt, die Zuweisung sowohl der GEN-Relation als auch einer PRP-Relation. So wird ENTSCHEIDUNG G KOMITEE Dokumenten zugeordnet, die z.B. die Phrase "die Entscheidung des Komitees" oder "die Entscheidung vom Komitee" enthalten. Umgekehrt wird EXTRAKT P ROSE auf Dokumente zugreifen, die z.B. "Extrakt der Rosen" oder "Extrakt aus Rosen" enthalten. Es erschien zweckmäßiger, den Ballast redundanter Formen in Kauf zunehmen, als vom Benutzer zu verlangen, mit EXTRAKT G ROSE bzw. ENTSCHEIDUNG P KOMITEE zu recherchieren. Zudem kann durch das Retrievalsystem eine Synonym-Relation zwischen beiden Varianten (beim erstmaligen Auftreten) erstellt werden.

Wenn das Analyse-Modul, auf das das Synthesesystem zur Deskriptorerstellung zugreift, eine vollständige (Tiefen-)Kasusanalyse lieferte, wäre die Synthese ggf. entsprechend zu modifizieren. Dies ist jedoch gegenwärtig nicht der Fall.

Ebenso muss von der Analysestrategie bestimmt werden, was als Präposition - bzw. genauer: als Präposition eines Präpositionalattributs zu einem Nomen - zu werten ist. In Fällen wie IN, AUS, DURCH, FÜR usw. ist es kein Problem.; Grenzfälle sind jedoch BETREFFEND usw., auch existiert eine gewisse Neigung zur 'Präpositionalität' bei manchen komparativen Adjektiven wie HÖHER ALS vs. ÜBER. In manchen Fällen wird die Frage, ob ein Wort den Status einer Präposition besitzt, durch die Analysestrategie entschieden, in anderen Fällen durch die Notwendigkeiten der Übersetzung bzw. der fachlichen Differenzierung, und schließlich durch die Aufnahme oder Nichtaufnahme ins morphologische Lexikon. Im konkreten Fall wurden die Entscheidungsmöglichkeiten für CTX von Überlegungen während der Entwicklung und Implementierung der Analysekomponente bestimmt.

KON-Relation

Die Anreihungsrelation (KON-Relation) wird zunächst zwischen angereihten Elementen nominaler Ausdrücke erkannt. So liefert die Phrase

"Farben und andere Chemikalien"

sowohl den Komplexen Deskriptor FARBE K CHEMIKALIE

als auch die Einfachen Deskriptoren FARBE
und CHEMIKALIE.

Die KON-Relation weist einige Merkmale auf, die sie von den anderen Typen Komplexer Deskriptoren unterscheidet. Erstens ist sie kommutativ: eine explizite Relation T1 K T2 impliziert T2 K T1, im obigen Beispiel CHEMIKALIE K FARBE. Zweitens ist sie assoziativ. Bei dem Beispiel "Farben, Pestizide, Herbizide" implizieren die Komplexen KON-Relationen FARBE K PESTIZID und PESTIZID K HERBIZID auch FARBE K HERBIZID sowie die Kommutation aller drei Paare. Mit anderen Worten liefert ein koordinierter Ausdruck mit n einfachen nominalen Deskriptoren (n-1) explizite und (n-1)(n-1) implizite, oder insgesamt n(n-1) Komplexe Deskriptoren. Der Zweck der Erkennung aller möglichen binären Kombinationen und Permutationen eines zusammengesetzten Ausdrucks liegt auf der Hand. Der Benutzer kann nicht im voraus den wörtlichen Inhalt der Dokumente, die er sucht, kennen; und irgendeiner der 6 komplexen Ausdrücke wird in dem Beispiel genügen, um das Dokument zu finden, in dem der oben genannte mehrwortige Ausdruck vorkommt. Die Frage stellt sich nun, wie diese Relationen gespeichert werden sollen.

(1) Explizite Relationen

Explizite Relationen müssen in der Deskriptorenliste enthalten sein, die dem betreffenden Dokument zugeordnet wird; d.h. in der Liste, auf die das IR-System während des Retrievalvorgangs zugreift. Jede andere Lösung würde das Durchblättern des Dokuments und eine syntaktische on-line-Textanalyse voraussetzen.

(2) Implizite Relationen

Theoretisch kann man kommutative und transitive Relationen aus den expliziten folgern, entweder während der Generierung von Deskriptorenlisten oder während des Retrievals. Trotzdem gibt es zwei Gründe, warum letzteres nicht praktikabel ist. Erstens muss angenommen werden, dass das IR-System, das CTX verwendet, keine Funktion zur Ausführung dieser Form der Analyse bereitstellt. Der Begriff des 'Komplexen Deskriptors' stellt ein besonderes Konzept des CTX-Verfahrens dar. In (herkömmlichen) IR-Systemen wird ein Komplexer Deskriptor als ein Buchstabenstring wie jeder andere Deskriptor behandelt (vgl. aber die Verfahrensweise von CONDOR, WIELAND 1979). Implizite Relationen müssen grundsätzlich nur einmal abgeleitet werden, wenn sie im Verlauf der Generierung der Deskriptorenkette bestimmt werden, im Retrievalvorgang jedoch so oft, wie sie jeweils angesprochen werden. Die geringen Einsparungen an Recherchezeit, die aus dem verminderten Umfang der Deskriptorenliste resultieren, würden mehr als ausgeglichen durch die erhöhte Rechenzeit, die durch real-time-Generierung impliziter Relationen entsteht. Zudem kann auch hier durch Generierung einer Synonymie-Relation im Thesaurus des IR-Systems der Speicheraufwand reduziert werden.

Es lässt sich entsprechend die Möglichkeit denken, implizite Relationen während des Retrievals mit einer natürlichsprachigen Problembeschreibung zu erkennen. In diesem Fall wird die Erkennung von einer Funktion ausgeführt, die Suchanfragen verarbeitet, bevor sie an das IR-System weitergegeben werden. Dabei kann wie folgt verfahren werden:

(a) Behandlung kommutativer Relationen:

Die Spezifizierung irgendeiner Komplexen Relation T1 K T2 durch den Benutzer schließt automatisch die kommutative Relation T2 K T1 in die Recherche ein: eine Recherche deckt daher jeden Komplexen Deskriptor mit einer K-Relation ab, der diese beiden Ausdrücke enthält - ungeachtet der Reihenfolge, in der sie in der Deskriptorenliste vorkommen. Dieses Vorgehen reduziert die Menge der zu speichernden KON-Relationen um die Hälfte, wobei die Effizienz des Retrievals nicht beeinträchtigt wird.

(b) Behandlung assoziativer Relationen:

Die Assoziativitätsregel kann in der Suchanfrage auch auf nicht-binäre Deskriptoren angewendet werden. So ergibt die Phrase "Farbe, Schutzmittel und Pestizide" in der Suchanfrage FARBE K SCHUTZMITTEL, SCHUTZMITTEL K PESTIZID und implizit FARBE K PESTIZID sowie die drei entsprechenden Kommutationen. Diese Operation reichte aus, um ein Dokument zu finden, das den Ausdruck "Farben, Pestizide und Herbizide" auf der Basis des expliziten (nicht-assoziativen) Komplexen Deskriptors FARBE K PESTIZID enthält. Wäre die Problembeschreibung "Farbe, Schutzmittel und Herbizid", würde dennoch das gleiche Dokument über den Komplexen Deskriptor FARBE K HERBIZID gefunden, der sowohl in der Rechercheformulierung als auch in der Deskriptorenliste assoziativ ist. Wie nützlich aber dieses Verfahren auch sein mag, so kann es doch nicht die Generierung assoziativer Relationen in der Deskriptorenliste ersetzen.

Die Verfahrensbausteine von CTX beinhalten die automatische Bereitstellung sowohl

assoziativer als auch kommutativer KON-Rela definiert und realisiert. Deren Hauptbestandteil ist das an der Oberfläche in verbalisierter Form auftretende Prädikat eines Satzes. Zum einen wird dabei die Relation zwischen einem Akkusativ-Objekt und dem zugehörigen Verb (AKK-VRB-Relation) ausgenützt, zum anderen die Beziehung zwischen einem Verb und dem ihm assoziierten Modalverb (MOD-VRB-Relation). So liefert z.B. der Satz

"Das Bundessprachenamt muss Daten speichern."

| | | |
|---|---|---|
| sowohl | DATEN SPEICHERN | (AKK-VRB-Relation) |
| als auch | SPEICHERN MUESSEN | (MOD-VRB-Relation). |

Während die Einbeziehung einer AKK-VRB-Relation von allgemeinerem Interesse ist, stellt die MOD-VRB-Relation ein Beispiel für die Möglichkeit einer Anpassung an fachspezifische Bedürfnisse dar. Die Laboranwendung des CTX-Verfahrens war im 2. Projektzeitraum (JUDO-DS) auf juristische Texte im Bereich Datenschutz ausgerichtet. Im fachlichen Interesse liegt es gelegentlich, ob in einem Dokument von "speichern können" oder "speichern müssen" die Rede ist. Für die praktischen Bedürfnisse ist diese Art der Präkoordination allerdings erst ein vorläufiger Schritt, da noch eine Reihe von zusätzlichen Schritten der Paraphrasierung erforderlich scheinen, etwa die (Quasi-) Synonymverknüpfung von NUTZUNGSRECHT und NUTZEN KOENNEN, von DATEN SPEICHERN und DATENSPEICHERUNG bzw. SPEICHERUNG G DATUM u.a.m. Hierzu werden im folgenden Abschnitt einige Ansätze geschildert.

### I.2.6 Das Problem der Paraphrase

Ein wesentliches Problem im Zusammenhang zwischen Indexierung und Retrieval ist die Möglichkeit, etwas mehr oder weniger gleich "Gemeintes" (oberflächen-)sprachlich unterschiedlich auszudrücken. Die moderne Linguistik hat inzwischen hierzu sowohl theoretisch wie empirisch wesentliche Erkenntnisse vermittelt. Von diesen Erkenntnissen ist bislang jedoch so gut wie nichts in die automatischen Verfahren zu Texterschließung und -retrieval übernommen worden. Im Gegenteil: diese Probleme gelten bislang als weitestgehend maschinell unlösbar, so dass Forschungen, die in diesen Bereich vordringen, auf sich alleine gestellt sind und zudem noch in den Beweiszwang der praktischen Realisierbarkeit geraten. Andererseits müssen notwendig Lösungsvorschläge zu diesem Themenbereich angeboten werden, will man nicht den Erfolg von Teilschritten - z.B. die Verwertbarkeit der Komplexen Deskriptoren - in Frage stellen.

Die im Projektzeitraum hierzu durchgeführten Forschungen mussten sich auf Teilbereiche beschränken. Von diesen Ergebnissen konnten wiederum nur Teilaspekte auch empirisch-technisch umgesetzt werden. Diese sollen hier aufgeführt werden, da damit einerseits die Bedeutung der Lösung des Paraphrase-Problems unterstrichen wird, andererseits die Ergebnisse zeigen, dass bereits mit einfachen Mitteln beträchtliche Fortschritte erzielt werden können, die v.a. die Integration der beschriebenen Präzisierungskomponenten, so z.B. die Bedeutungsdifferenzierung und die Vergabe von Komplexen Deskriptoren, rechtfertigen.

#### I.2.6.1 Lexikalische Synonymie und Begriffsvernetzung

In einem deskriptororientierten System ist die Integration der lexikalischen Synonymie von grundlegender Bedeutung. Ihr wird im Rahmen eines begrifflichen Relationenlexikons Rechnung getragen (CTX-Thesaurus). Zu den (weiter differenzierten) Synonym-Relationen rechnen einmal:

- die "strenge" Synonymie
(z.B. FAHRSTUHL (SYN) AUFZUG)

- die Relation Langform-Abkürzung
(z.B. BUNDESDATENSCHUTZGESETZ (LNG) BDSG)

- die Schreibvarianten-Synonymie
(z.B. NIESSBRAUCHRECHT (SYS) NIESSBRAUCHSRECHT)

- die (verschiedenen) Derivationsrelationen
(z.B. WOHNEN-WOHNUNG-WOHNLICH

Zusätzlich zu den üblichen semantischen Synonym-Relationen beinhaltet der CTX-Thesaurus also so genannte "Derivationen" zwischen Substantiven, Verben und Adjektiven. So führen die drei Begriffe NACHWEIS, NACHWEISEN, NACHWEISLICH zu den folgenden Relationen:

| | | |
|---|---|---|
| NACHWEIS | (DSV) | NACHWEISEN |
| NACHWEIS | (DSA) | NACHWEISLICH |
| NACHWEISEN | (DVS) | NACHWEIS |
| NACHWEISEN | (DVA) | NACHWEISLICH |
| NACHWEISLICH | (DAS) | NACHWEIS |
| NACHWEISLICH | (DAV) | NACHWEISEN |

Der Zweck von Derivationsrelationen liegt in der Verbesserung der "Reichweite" eines gegebenen Suchbegriffs und damit des Recalls - und nicht darin, Informationen über die morphosyntaktischen Beziehung zwischen Begriffen bereitzustellen. Die Relation ist also in erster Linie semantischer Natur (d.h. eine Art Synonymie) und erst in zweiter Linie syntaktischer. Dies bedeutet zweierlei: Erstens wird nicht versucht, die historischen Gegebenheiten der Wortbildung zu beschreiben, so dass jede Derivationsrelation T1 (D) T2 weder die historische noch eine direkte morphologische Ableitung bedeuten soll. Es kann also historische Zwischenstufen in der Derivation gegeben haben, wie z.B. in NATION - NATIONAL - NATIONALISIEREN. Folglich sind diese Relationen einerseits kommutativ: T1 (Dxy) T2 beinhaltet T2 (Dyx) T1, andererseits transitiv: T1 (Dxy) T2 und T2 (Dyz) T3 beinhaltet auch T1 (Dxz) T3.
Zweitens werden Derivationen dort nicht erstellt, wo die morphologische Derivation von einer Verschiebung des semantischen Feldes begleitet ist, d.h. wo zwei Begriffe für die Suchanfrage nicht wechselseitig bedeutsam sind, z.B.

| | | | |
|---|---|---|---|
| | KREIEREN | (DVA) | KREATIV |
| und auch | KREATION | (DSA) | KREATIV |
| aber nicht | KREATUR | *(DSA) | KREATIV. |

Derivationsrelationen zwischen Wörtern gleicher Wortklasse (DSS-, DVV- oder DAA-Relationen) werden nur in Ausnahmefällen verwendet: Erstens beinhalten wortklassengleiche Derivationen im typischen Fall semantische Verschiebungen:

| | | |
|---|---|---|
| FORM | *(DSS) | FORMATION. |

Zweitens gibt es für den Ausnahmefall, in dem die semantischen Felder zueinander passen, bereits genügend Möglichkeiten, um diese Relation zu speichern:

FREIWILLIG (GEG) UNFREIWILLIG

Eine Schwierigkeit stellen die Homonyme dar. In einigen Fällen kann eine Gruppe von Homonymen einer syntaktischen Klasse einer derivationsrelationierten Gruppe einer anderen Klasse entsprechen. In anderen Fällen jedoch ist die Homographie in den Derivationen nicht konsistent. Dann kann nicht von einer Derivationsrelation gesprochen werden.

Eine weitere Schwierigkeit bedeuten Nominalisierungen, die produktivsten aller Derivationen im CTX-Thesaurus. Bei jedem Verb v, aus dem ein deriviertes Nomen n gebildet werden kann, hat das Nomen im allgemeinen zwei Bedeutungen:

nl: die Handlung des v-s (immer im Singular)
n2: das Ergebnis des v-s (im Singular oder im Plural)

z.B.: ILLUSTRATION

(1) die Handlung des Illustrierens:
die Illustration eines Vorgangs
(2). etwas, das illustriert:
Dieses-Buch enthält viele Illustrationen.

Wenn diese beiden Bedeutungsvarianten im Thesaurus differenziert sind, muss die Art der DVS-Relationen näher definiert werden. Hierfür gibt es zwei Möglichkeiten. Die erste ist die Verwendung von DVS- und DSV-Relationen zur Markierung der semantischen Ähnlichkeit (Synonymie), wobei die obige Unterscheidung vernachlässigt wird; beide Möglichkeiten sind daher zuzulassen: ILLUSTRIEREN (DVS) ILLUSTRATION1 und ILLUSTRIEREN (DVS) ILLUSTRATION2 (und ihre Invertierungen). Diese Lösung besitzt den Vorteil, mit der Verwendung der übrigen Derivationsrelationen zu harmonieren.

Eine zweite Möglichkeit ist, die DVS- und DSV-Relation spezifischer zu verwenden, um die Relation zwischen einem Verb und seiner Nominalisierung darzustellen. Dabei wird eine Beziehung zwischen Texten bewahrt, in denen Vorgänge explizit als Verben dargestellt werden, und anderen, in denen diese Vorgänge aus stilistischen Gründen in nominalisierter Form erscheinen - dies ist z.B. ein Merkmal formeller und technischer Schriften im Gegensatz zu journalistischen Berichten und Gebrauchsliteratur. In diesem Fall würde nur ILLUSTRIEREN (DVS) ILLUSTRATION1 bestimmt. Diese Lösung hat den zusätzlichen Vorteil, den Unterschied in der Zugriffsmöglichkeit auf diese Nominalisierungen und ähnliche syntaktische Konstruktionen, in denen Partizipien verwendet (und daher schon in der syntaktischen Analyse als Verben erkannt) werden, zu beseitigen.

Der Nachteil dieser zweiten Möglichkeit ist, dass DSV- und DVS-Relationen nicht zusätzlich verwendet werden können, um Dokumente auf rein semantischer Grundlage zu finden. Eine mögliche Lösung hierfür wäre die Einführung eines besonderen Relationstyps, um Vorgangs-Nominalisierungspaare (z.B. DVN und DNV) zu markieren, wobei DVS und DSV weiterhin ihre mehr semantische Funktion ausüben könnten. Also könnte der Thesaurus folgende Derivationen

enthalten:

| | |
|---|---|
| ILLUSTRIEREN | (DVN) ILLUSTRATION1 |
| ILLUSTRIEREN | (DVS) ILLUSTRATION1 |
| ILLUSTRIEREN | (DVS) ILLUSTRATION2 |
| ILLUSTRATION1 | (DNV) ILLUSTRIEREN |
| ILLUSTRATION1 | (DSV) ILLUSTRIEREN |
| ILLUSTRATION2 | (DSV) ILLUSTRIEREN |

Der Benutzer könnte dann entweder nach stilistischen (mittels DVN und DNV) oder nach semantischen Äquivalenten (mittels DSV und DVS) suchen (d.h. in der Praxis: entsprechende Parametrisierungen beim Retrieval bewirken).

Im weiteren Sinne lassen sich auch die Quasi-Synonymie und die Assoziation zu diesem Problemkreis rechnen.

### I.2.6.2 Halblexikalische Synonymie

Vor allem in deutschsprachigen Texten muss man mit vielen sog. "Augenblickskomposita" rechnen, d.h. Phrasen, die in Texten einmal als Kompositum (d.h. zusammengesetztes Wort), einmal als Mehrwortbegriff auftreten können. So kann der Benutzer eines IR-Systems etwa die Formel BUNDESKANZLERREISE verwenden, während im Text (Dokument) die Wendung REISE DES BUNDESKANZLERS vorkommt.

Im implementierten Verfahren wird diesem Phänomen durch die Aufnahme des Mehrwortbegriffs als Synonym zu einem lexikalisierten, d.h. im fachgebietsspezifischen Lexikon erfassten Kompositum in den CTX-Thesaurus Rechnung getragen; die "Augenblickskomposita" werden dabei auch mit ihren lexikalischen Bestandteilen verzeichnet.

Hierzu lassen sich jedoch noch weitere Verfahren denken, etwa die automatische Generierung einer P- bzw. G-Relation zu einem Augenblickskompositum und umgekehrt die "künstliche" Komposition bei Komplexen Deskriptoren, etwa - wie schon erwähnt - die Nominalisierung von AKK-VRB-Relationen:

DATEN SPEICHERN - DATENSPEICHERUNG
SPEICHERN MÜSSEN - SPEICHERUNGSPFLICHT.

Dies soll ggf. in einer späteren Ausbaustufe des Systems CTX bedacht werden .

### I.2.6.3 Syntaktische Paraphrasen

Beispiele für syntaktische Paraphrasen wurden bereits in den vorangegangenen Abschnitten aufgeführt. Während ein Teil davon (Z.B. die Aktiv-Passiv-Struktur) bereits in dem transformationellen Teil des verwendeten automatischen Sprachanalysesystems berücksichtigt wird, und damit bei der Indexierung (etwa der Erzeugung der AKK-VRB-Relation) also außer Acht bleiben kann, sind noch einige andere Verfahrensschritte zu implementieren (etwa die Verknüpfung der Relation Adjektiv-Nomen mit der Prädikativform (PERSONENBEZOGENEN DATEN - DATEN heißen/sind PERSONENBEZOGEN, wenn ...), insbesondere auch unter Berücksichtigung der relati-

vischen Anschlüsse).

### I.2.6.4 Grenzen der syntakto-semantischen Paraphrasierung

Setzt man voraus, dass die beschriebenen Fälle von Paraphrasierung in einem IR-System angemessen berücksichtigt sind, so ist ein entscheidender Schritt in Richtung auf eine benutzerfreundliche Textdokumentation getan. Auf dieser Grundlage wird v.a. der Präzisierungskomponente voll Rechnung getragen, wie sie mit dem Prinzip der Komplexen Deskriptoren intendiert ist.

Dennoch bleiben - dies soll nicht verschwiegen werden - auch auf dieser Ebene eine Reihe von Fragen offen. Da das System z.B. keine Handlungszusammenhänge erkennt bzw. kennt, können z.B. Fakten, die in einer Aussage implizit enthalten sind, nicht unmittelbar verfügbar gemacht werden. Dafür ein Beispiel: In einem Gesetz stehe etwa "Der Betroffene ist berechtigt, eine kostenlose Anfrage an die speichernde Stelle zu richten". Versucht es der Informationssuchende nun mit den Begriffen "BRIEF" und "SCHREIBEN", so wird er nicht auf dieses Dokument verwiesen. Umgekehrt gilt, dass jemand, der mit dem Begriff "ANFRAGE RICHTEN" sucht, wohl dieses Dokument, nicht aber ein anderes findet, in dem vielleicht von dem "Schreiben eines formlosen Briefes" die Rede ist. In beiden Fällen ist letztlich nur über komplexere Verfahrensweisen Abhilfe zu schaffen. Ansatzweise - ohne Anspruch auf Vollständigkeit - wurde im Rahmen des Forschungsprojekts der CTX-Thesaurus zum Bereich Datenschutzrecht auf diese Problematik hin ausgerichtet. Im Rahmen der verfügbaren hierarchischen und assoziativen Thesaurusrelationen lässt sich nämlich derartiges "Weltwissen" wenigstens partiell integrieren. Allerdings wird damit ein Bereich angesprochen, der den Verfahrensweisen der Systeme der Künstlichen Intelligenz ähnelt, ohne jedoch deren Präzisions- und Formalisierungsgrad zu erreichen. Die geringere Präzision (d.h. die größere Vagheit) kann jedoch auch von Vorteil sein. Zumindest ist der intellektuelle Kodieraufwand noch überschaubar. Es scheint jedoch, dass diese "Weltwissen"-bezogenen Relationierungen zusätzlich noch fachgebiets- und themenspezifisch differenziert werden müssen, will man den Benutzer nicht durch die (Über-)Fülle von möglichen Vernetzungen beim Retrieval ratlos werden lassen.

## I.2.7 Die Schnittstelle zum Information-Retrieval-System

Das Hauptanliegen der Forschungen konnte zunächst ohne Bezugnahme auf irgendein spezielles IR-Software-System beschrieben werden. Zielsetzung war es, aufgrund einer syntakto-semantischen Analyse eines Textes Einfache und Komplexe Deskriptoren als formal-inhaltliche Repräsentanten eines Text-Dokuments zu erkennen und zu bestimmen. Für das Retrieval werden dabei folgende Teile dem System verfügbar gemacht:

(1) der Text des Dokuments
(2) die Deskriptoren
(3) neue Einträge bzw. Vorschläge in Form von Ausdrücken und Relationen für den IR-Thesaurus
(4) einige weitere Einzelheiten der Analyse, einschließlich der erkannten Struktur und der grammatischen Merkmale, die den strukturellen Elementen zugewiesen werden.

Diese Informationen müssen den vorgegebenen Konventionen der benutzten IR-Systeme (Hosts) angepasst werden. Zu diesen Fremdsystem-eigenen Konventionen gehören v.a. Kommandoworte, Datenformate und Ablauf-Konventionen. Dabei sieht CTX eine Schnittstelle zu möglichst vielen

potentiell anwendbaren IR-Systemen vor. Die ggf. erforderlichen Schnittstellenprogramme werden als Datenumsetzungssystem (DUMS) bezeichnet. Zwei Schnittstellen-Versionen wurden im Projektverlauf zur empirischen Erprobung der Ergebnisse implementiert: DUMS-T und DUMS-G, entsprechend den beiden IR-Systemen TELDOK und GOLEM, an die das Modell-System bisher angeschlossen wurde.

Offensichtlich gibt es einige Minimalanforderungen an jedes in Betracht kommende IR-System. Ungeachtet technischer Überlegungen, wie der Verfügbarkeit von Hardware und ausreichender Speicherkapazität, muss das IR-System mindestens die Speicherung der Texte ermöglichen, ferner die Verwaltung einer beliebigen Menge von Deskriptoren, die diesem Text zugewiesen sind, sowie Retrievalfunktionen zur Ermittlung von Textteilen ("Dokumenten") mithilfe der zugeordneten Deskriptoren. Neben diesen grundsätzlichen Aspekten bieten alle anderen von einem IR-System bereitgestellten Faktoren ein Leistungsspektrum, das im Rahmen des Forschungsprojekts in mehrerlei Hinsicht benutzt wurde, um den Informationsverlust zwischen CTX-System und IR-Benutzer zu verringern. Eine solche Benutzung musste gelegentlich Funktionen in ganz anderer Weise anwenden, als es eigentlich seitens des IR-Rahmensystems vorgesehen war. Die Folge ist, dass bei der Datenumsetzung gelegentlich auf "Hintertür"-Techniken zurückgegriffen wurde, um seitens des IR-Systems nicht unmittelbar bereitgestellte Funktionen zu realisieren. Art und Ausmaß, in denen die Ergebnisse von CTX - über DUMS - durch das fremde IR-System eingeschränkt werden, behandeln die folgenden Abschnitte.

### I.2.7.1 Behandlung des Dokumenttextes

Der Text in seiner ursprünglichen Form (Original) muss sowohl für die Analyse-Software als auch für die speichernde IR-Software akzeptabel sein (in den Pilotanwendungen also für TELDOK und GOLEM). Der Originaltext muss derzeit über zwei Schnittstellen eingegeben werden:

(1) Analyse-Schnittstelle.

Die Eingabe für die Analyse setzt einige Umformatierungen und damit eine gewisse Editierung (Prä-, Post- oder Online-) voraus. Das Saarbrücker Übersetzungsystem beispielsweise nimmt Texte gegenwärtig nur in bestimmter formaler Aufbereitung an. Der Satzendepunkt ist z.B. durch einen Asterisk zu ersetzen; besondere Symbole sind eingeführt, um Überschrifts- und Abschnittsende oder andere Textmarkierungen zu kennzeichnen. Ein Großteil dieser erforderlichen Transkriptionen aus "normalen", d.h. am Ausgangstext orientierten Texterfassungen kann - wenn maschinenlesbar - vollautomatisch erfolgen; ein anderer Teil muss intellektuell geschehen, wenn auch ggf. mit maschineller Unterstützung (z.B. bei der Erkennung des Satzendepunktes). Ein Teil erfolgt halbautomatisch über eine intellektuelle Bewertung der vom Rechner generierten Ergebnisse (z.B. Bestätigung jeder erkannten Markierung eines Abschnittsendes).

(2) Speicherschnittstelle:

Der Text kann prinzipiell zur Speicherung im IR-System (über das Datenumsetzungssystem) entweder vom Original oder vom Output der Analyseschnittstelle her aufbereitet werden. Es gibt jedoch einige praktische Gründe dafür, die erste Lösung vorzuziehen. Erstens dienen die Textänderungen, die für die Analyse vorgenommen werden, vorwie-

gend der maschinellen linguistischen Analyse: diese zusätzlichen Informationen für die Analyse sind im allgemeinen für Datenbank-Speicherzwecke irrelevant; andererseits sind einige Informationen, die sich auf das Layout des Dokuments beziehen, überflüssig für Analysezwecke und würden (evtl. mangels geeigneter Zwischenspeicherungsmöglichkeiten: endgültig) durch die Analyse-Outputschnittstelle getilgt. Zweitens würden Speicherschnittstellen, die von einer bestimmten Analyseschnittstelle abhängen, ein hohes Maß an unerwünschter Unbeweglichkeit im ganzen System zur Folge haben: DUMS-"X" (für ein beliebiges IR-System "X") müsste für jedes in CTX integrierte Analyse-System umgeschrieben werden (z.B. für verschiedene Sprachen, verschiedene thematische Bereiche). Drittens kann ein IR-System sehr wohl sein eigenes Editierungsverfahren haben (das von DUMS-X ausgenutzt werden könnte), dessen Eingabe vermutlich ein normaler Text ist und nicht der präedierte Analyse-Input. Daher bietet es sich an, den gleichen Text, der ursprünglich als Input für die Analyseschnittstelle diente, an das Datenumsetzungssystem weiterzugeben, das seinerseits den Text entsprechend den Anforderungen des fremden IRSystems in das spezifische Speicherformat bringt.

### I.2.7.2 Deskriptorenlisten

Deskriptorenlisten sind nicht so stark strukturiert wie die Texte selbst; Einzelheiten wie Interpunktionen, Groß-/Kleinschreibung, Unterstreichung, Zeichenvorrat sowie die Darstellung von Zahlen und anderen nicht "echten" Wörtern sind bis auf einige Zeichenkonventionen unerheblich. Deshalb ist die Formatierung von Deskriptoren wesentlich einfacher; willkürliche Unterschiede der Darstellung zwischen Deskriptoren des CTX-Systems und des Host-IR-Systems sind wahrscheinlich trivial. In diesem Fall ist die Schnittstelle unproblematisch (und vollautomatisch realisierbar). Das völlige Fehlen einer internen Struktur von Deskriptoren bedeutet jedoch für die Datenumsetzung zwei Schwierigkeiten, eine mögliche und eine tatsächliche. Die potentielle Schwierigkeit hängt mit der Darstellung von Homographen und Homonymen zusammen. Da jeder Begriff im Thesaurus eindeutig sein muss, müssen alle (natürlichsprachigen) Begriffe in diesem Thesaurus, die die gleiche "Oberflächendarstellung" (Buchstabenkette) aufweisen, irgendwie unterschieden werden (d.h. zu "kunstsprachigen Begriffen" gemacht werden). Gewöhnlich stellt das keine besondere Schwierigkeit dar. Die meisten IR-Systeme akzeptieren Ziffern als vollwertige Buchstaben, und so liegt die Lösung auf der Hand, nämlich einen Homographen bzw. ein Homonym mit zwei Bedeutungen als Variante 1 und Variante 2 zu bezeichnen (unter Zulassung auch einer unmarkierten Variante); dies ist bei den intern generierten Deskriptoren von CTX auch tatsächlich der Fall. Wenn das IR-System Ziffern zurückweist, sind andere Lösungen verfügbar (z.B. Darstellung als Variante A oder Variante B, oder auch andere Vereindeutigungen, wie z.B. Paraphrasierungen bzw. Definitionen als Übernahme aus dem entsprechenden Lexikon). Dies war jedoch weder bei TELDOK noch bei GOLEM notwendig.

Für normale IR-Zwecke ist diese Lösung vollkommen ausreichend: Auf eine Thesaurus-Anfrage hin wird ausgegeben, dass Variante 1 und Variante 2 deskriptorfähige Begriffe sind, und der Benutzer kann seine Suchbegriffe dementsprechend formulieren. Praktische Probleme entstehen bei abhängiger Software - z.B. bei der Erstellung einer sortierten Liste von Stichworten. Wenn Ziffern hinter den Buchstaben A - Z sortieren, wird die sortierte Liste eine Reihenfolge aufweisen, bei welcher der Homographenstamm möglicherweise (z.B. durch Wortzusammensetzungen) von den Deskriptor-Varianten 1 und 2 getrennt wird. Lösungen für dieses Problem sind bekannt, aber es ist unwahrscheinlich, dass sie in ein IR-Software-System integriert werden, für das die Verwendung von Techniken zur Auflösung von Homographen in dieser Form nicht vorgesehen ist.

Wenn sortierte Listen (entweder hardcopy oder in der Art einer on-line-Recherche im Thesaurus) einen Teil der Suchverfahren im IR-System bilden, betrifft dieses Problem den Benutzer und nicht nur den Archivar oder Systemverwalter.

Das tatsächliche Problem stellt sich in Verbindung mit dem Konzept des Komplexen Deskriptors (s.o.). Da es hier keine expliziten Regeln zur Erkennung von "Teilen" von Deskriptoren gibt (außer einzelnen Buchstaben zum Pattern-Matching (s. Teil I.), haben Komplexe Deskriptoren (einschließlich ihrer Relatoren P, G, K) für das IR-System das gleiche "Aussehen" wie alle anderen mehrwortigen Deskriptoren. Einige Schwierigkeiten, die so entstehen, werden gegenwärtig in der Anwendung des Modellsystems dadurch umgangen, dass sowohl die Bestandteile, d.h. die betreffenden Einfachen Deskriptoren, als auch der Komplexe Deskriptor als Deskriptor einem Dokument zugeordnet sind, so dass der Benutzer über Einzelbestandteile als auch komplexe Ausdrücke suchen kann - wenn auch auf Kosten einer gewissen Redundanz in den Dokument-Deskriptoren. Zusätzlich sorgt eine spezielle Relation, vergeben als Thesaurus- Beziehung, dafür, dass alle zu einem Einzeldeskriptor vergebenen Komplexen Deskriptoren angezeigt bzw. in die Suchfrage integriert werden können.

### I.2.7.3. Thesaurus-Einträge

Deskriptoren, die in einem neu erfassten Text zum ersten Male auftreten, müssen in den IR-Thesaurus mit allen neuen Relationen zu den bereits vorhandenen Einträgen aufgenommen werden. Dies geschieht unter Verwendung des CTX-Thesaurus, der dem Thesaurus des IR-Systems "logisch" entspricht. Die realisierten automatischen halbautomatischen Thesaurus-Verfahren stellen gegenwärtig eher Funktionen des CTX-Thesauruspflegesystems denn solche eines "fremden" IR-Systems dar.

Die Implementierung der TELDOK-Variante hat gezeigt, dass in der praktischen Umsetzung Informationen zwischen dem CTX-Thesaurus und dem Thesaurus des IR-Systems verloren gehen können. Eine der Beschränkungen des TELDOK-Systems war, dass Thesaurus-Relationen sowohl in der Anzahl von Typen (max. 6) als auch in der Art begrenzt sind. TELDOK "kannte" systemseitig eine Differenzierung nach

- Synonymie
- Antonymie
- Hypernymie
- Hyponymie
- semantisches Feld
- Homonymie.

Die Homonymie wurde bei der Modellanwendung (hierbei der Logik des IR-Systems entsprechend) zum Hinweis auf die Mehrdeutigkeit der Zeichenkette verwendet. Hierzu waren im Hostsystem entsprechende Hinweisfunktionen implementiert. Auf die verbleibenden fünf Relationen mussten die vielfältigen Relationen des CTX-Thesaurus angepasst ("abgebildet") werden. Dies erforderte die Zusammenfassung der CTX-Relationen auf TELDOK-Relationen.

| CTX-Thesaurus | TELDOK-Thesaurus |
|---|---|
| Synonym/Langform/ | |

| | |
|---|---|
| Kurzform/Schreibvariante/ Derivation | Synonym |
| Antonym | Antonym |
| Hypernym/Ganzes/Gruppe | Hypernym |
| Hyponym/Teil/Glied | Hyponym |
| Quasi-Synonym/Assoziation | Semantisches Feld |

Ungeachtet der Verwendung in einem IR-Host-System wird also eine beträchtliche Anzahl von Relationen im CTX-eigenen Thesaurus differenziert. Dadurch wird nicht nur eine Unabhängigkeit von potentiellen Hostsystemen gewährleistet, sondern auch die Konsistenz des Systems durch Anwendung der Gesetze von Kommutativität, Transitivität und Assoziativität innerhalb des Thesaurus gewahrt.

Im Gegensatz zu TELDOK erlaubt das zweite modellhaft verwendete IR-Host-System GOLEM den Aufbau von bis zu 127 Relationstypen - mehr als genug für die Zwecke der Modellanwendung. Von diesen Relationen ist systemseitig nur eine Synonym-Relation vordefiniert. GOLEM führt dabei eine automatische Kommutierung von Synonymen durch, d.h. T1 SYN T2 erzeugt auch T2 SYN T1. Da CTX die Kommutierung im systemeigenen Thesaurus-Verwaltungssystem bereits vornimmt, analog zu den anderen (Transitivitäts- und Assoziativitäts-)Operationen, die seitens des GOLEM-Systems nicht realisiert sind, ergibt sich hierdurch eine partielle systematische Redundanz.

Der Vorteil einer größeren Auswahl von Relationen ist, dass die Suche feiner differenziert werden kann: Ein Nachteil kann darin liegen, dass der Benutzer gezwungen ist, komplexere Formulierungen zu verwenden (und damit, sich zuvor tiefere Systemkenntnisse anzueignen), wenn sich seine Suche auf allgemeinere Relationskategorien bezieht. So kann die Dokumentmenge, die der Benutzer des TELDOK-Systems mit der Einbeziehung der Relation "Semantisches Feld" erhält, in der GOLEM-Version durch Einbeziehung (entweder jeweils explizit oder durch Voreinstellung) einer ganzen Reihe von Relationen ermittelt werden: QUA, ASS, REG, IUS und NEB. Im Idealfall sollte der Benutzer daher nicht nur einen Sammelbegriff spezifizieren können, sondern auch den Grad an Allgemeinheit, den er bei seiner Thesaurusanfrage wünscht - am besten mithilfe eines geeigneten einfachen Formalismus.

Die günstigste Lösung wären hierarchisch strukturierte Relationen, d.h. ein Definitionsschema, in dem Meta-Relationen zwischen den Relations-Kategorien selbst definiert werden könnten. In einem solchen Schema könnte der Benutzer alle äquivalenten, vertikalen oder fachgebietsspezifischen Relationen erfragen, außerdem alle Synonyme, Teile oder Hypernyme. Diese Hierarchie ist nicht unbedingt vollständig: Man könnte sich andere Meta-Gruppen vorstellen, wie z.B. eine Kategorie TANGENT, die HYPER und HYPO umfasst, oder eine Kategorie SYSTEM für GANZES und TEIL, oder ALTERNATIV für LANGFORM, KURZFORM und ORTHOGRAPHISCHE VARIANTE. Die KURZFORM könnte man weiterhin in AKRONYM und ABKÜRZUNG unterteilen oder entsprechend LANGFORM in VOLLFORM und NICHTABKÜRZUNG.

In GOLEM ist keine explizite Möglichkeit zur Strukturierung von Relationen (d.h. zum Aufbau und der ggf. voreingestellten Anwendung von Meta-Relationen) vorgesehen (uns ist überhaupt kein entsprechend ausgelegtes System bekannt). Man kann jedoch eine solche Hierarchie simulieren, indem man zusätzliche (Meta)-Kategorien definiert und innerhalb von CTX vielfache Relationen generiert. Hätte man dies über die gesamte Menge vorgenommen, wäre das Ergebnis

eine mehr als dreimal so große Zahl an Relationen im GOLEM-Wörterbuch als im CTX-Thesaurus gewesen. Je nach dem Grad der Verknüpfung in einer Hierarchie von willkürlicher Komplexität könnte der Expansionsfaktor von CTX zu GOLEM noch erheblich größer sein. Natürlich kann ein Thesaurus von ohnehin schon beträchtlicher Größe nur bis zu einer noch praktikablen Grenze "aufgebläht" werden: Ein IR-System, das theoretisch für die relativ niedrige Sättigung konzeptueller oder auf Wissen basierender Deskriptoren entwickelt worden ist, bedingt wahrscheinlich an irgendeinem Punkt hinter einem angenommenen "sicheren" Maximalwert physikalische oder operationale Beschränkungen. Dementsprechend nutzt die CTX-Implementierung dieses Potential nur beschränkt, Z.B. im Fall von Langform- oder Kurzformrelationen. Für jede solche Relation R bewirkt ein Relationspaar TI R T2 bei GOLEM die Generierung T1 R T2 und TI SYN T2, was zu der impliziten Hierarchie

SYN

SYN LANG KURZ ... (DERIVATIONEN)

führt. SYN ist dabei mehrdeutig; es identifiziert sowohl die Meta-Gruppe als auch die Meta-Glieder (worauf dabei nicht unabhängig zugegriffen werden kann). Dieser Kompromiss ist aufgrund der Häufigkeit synonymer Relationen im CTX-Thesaurus unumgänglich, und mit dem relativ starken Anwachsen des GOLEM-Wörterbuchs verbunden; dies wäre zwangsläufig der Fall, wenn jede SYN-Relation (einschließlich aller Invertierungen und Assoziationen) verdoppelt würde. Diese Hierarchie bewirkt, dass ein Benutzer während der Recherche die Suche mit einem Deskriptor ausdehnen kann: durch Einbeziehung der Langformen, der Kurzformen und der Synonyme. Die "strengen" Synonyme selbst können durch die Verwendung von Ausdrücken mit komplexerem Muster gewonnen werden, die explizit nach

SYN (X) UND NICHT (LANG (X) ODER KURZ (X))

fragen.

### I.2.7.4 Bereitstellung von Strukturdaten aus der Analyse

Im Hinblick auf Dokumenttext, Deskriptoren und Thesauruseinträge ist anzunehmen, dass das fremde IR-System zweckorientierte Verfahren für die Behandlung der Informationen bereitstellt, die von CTX geliefert werden. Auch wenn die Vollständigkeit dieser Informationen durch Beschränkungen des Fremdsystems stark beeinträchtigt werden mag, bleibt sie in Form und Inhalt im Wesentlichen doch unversehrt. Für die Speicherung von detaillierten Strukturdaten als Ergebnis der Textanalyse sehen gegenwärtige IR-Systeme - sieht man einmal ab von dem bei SIEMENS entwickelten Laborsystem CONDOR, das Ergebnisse der CONDOR-eigenen Sprachanalyse beim Retrieval verarbeiten kann - derzeit keine speziellen Funktionen vor, die etwa während des Retrievals zur Aufbereitung der erkannten grammatikalischen Struktur und anderer Merkmale, die mit dieser Struktur verbunden sind, verwendet werden könnten. Hier hängt die Datenumsetzung weniger von Beschränkungen bei der Implementierung von Standard-Verfahren eines potentiellen Fremdsystems ab; eine Abhängigkeit besteht vielmehr vom zufälligen Vorhandensein oder Fehlen spezieller Verfahren, die ein CTX-Retrievalbaustein für seine eigenen Zwecke einsetzen könnte - und die mit den Zielsetzungen während der Konzeption und Entwicklung des IR-Systems selbst nicht unbedingt übereinstimmen müssen.

Diese Sachlage ändert sich in Abhängigkeit vom jeweiligen Fremdsystem beträchtlich. Das System TELDOK bereitete die geringsten Schwierigkeiten, da es über verwertbare Hilfsmittel dieser Art gar nicht erst verfügte: Strukturelle Ergebnisse der Analyse können somit ganz einfach nicht in dem IR-System gespeichert und verwertet werden, so dass entsprechende Umsetzungsarbeiten entfallen. Der daraus resultierende Informationsverlust könnte als bedauerlich angesehen werden, ist aber für die Konzeption von CTX ohne Tragweite: die linguistische Analyse ist bereits ein notwendiger Schritt zur Erstellung von Einfachen und Komplexen Deskriptoren. Weitere (Zwischen-)Ergebnisse dieser Analyse werden - angepasst an die gegenwärtige Gesamtsituation in der Praxis marktgängiger IR-Systeme - eher als Nebenprodukt dieses Vorgangs gespeichert; jedes Hilfsmittel beim Retrieval, das diese Informationen ausnutzt, bedeutet einen Vorteil. Allerdings mindert das Fehlen solcher Hilfsmittel zu einem gewissen Grade die Rechtfertigung des rechnertechnischen Mehraufwands, den die linguistische Analyse erforderlich macht.

Erste Möglichkeiten einer zusätzlichen Informationsspeicherung und -nutzung bot dagegen die Implementierung auf dem GOLEM-Host-System. Der Datenbankverwalter hat die Möglichkeit, den Deskriptoren "Indices" und "Rollenindikatoren" zuzuweisen, die dann während des Retrievals zur sog. "Feinrecherche" zur Verfügung stehen. Jeder Index besteht aus einer Indexnummer; jeder Rollenindikator ist eine Folge von Zeichen, die frei vereinbart werden können; ihm kann wahlweise eine "Indexnummer" zwischen 1 und 255 vorausgehen. Rollenindikatoren, Indexnummern und trennende Kommas bilden zusammen einen String alphanumerischer Zeichen, der jeweils nicht länger als 255 Zeichen sein darf.

Die Indexziffern stellen eine Verbindung zwischen zwei oder mehr Deskriptoren her, wie sie durch die zugehörigen Rollenindikatoren definiert ist. Auf der untersten syntaktischen Ebene (d.h. bei den Einzelbegriffen) sind diese Verknüpfungen weitgehend identisch mit den Verknüpfungen, die in den Komplexen Deskriptoren enthalten sind. So würde z.B. die Phrase "Schutz vor Missbrauch" folgendes generieren

SCHUTZ ...121,PL
MISSBRAUCH ...121,PR
SCHUTZ P MISSBRAUCH ...

Dabei zeigt der gemeinsame Index-Wert (hier 121) an, dass SCHUTZ und MISSBRAUCH strukturell verknüpft sind, PL und PR markieren das linke bzw. rechte Element einer PRP-Relation. In der Feinrecherchephase zu einer Suchanfrage kann der Benutzer somit alle Dokumente (einschließlich des obigen Beispiels) finden, in denen SCHUTZ und MISSBRAUCH als linke und rechte Elemente vorkommen, bzw. als Elemente einer PRP-Relation. Dabei wird folgende Anfrage formuliert:

SCHUTZ *PL* MISSBRAUCH *PR*

In diesem Fall überprüft das System die Deskriptorenliste jedes gespeicherten Dokuments, das vorher durch eine "normale" Recherche (die in diesem Zusammenhang als "Grobrecherche" verstanden wird) vorausgewählt wurde, nicht nur auf das gemeinsame Auftreten von SCHUTZ und MISSBRAUCH, sondern zusätzlich darauf, ob im Deskriptorindex von SCHUTZ der Rollenindikator PL und im Deskriptorindex von MISSBRAUCH der Rollenindikator PR auftritt. Ist dies der Fall und stimmen die Indexziffern vor diesen Rollenindikatoren überein, so war die Suche erfolgreich und das Dokument wird als Suchtreffer ausgewiesen.

Dieses Verfahren stellt mehr als eine Alternative zum Vorgehen unter Verwendung des Komplexen Deskriptors dar: Über den Index können auch andere als terminale Strukturen abgebildet (und dementsprechend abgefragt) werden, z.B. die Zugehörigkeit eines Deskriptors zu Nominalgruppen, Teilsätzen oder Sätzen.

Umgekehrt ist der Komplexe Deskriptor das Ergebnis eines systeminternen Bausteins von CTX. Dieses entsteht unabhängig von der Fähigkeit irgendeines besonderen IR-Systems, explizite syntaktische Informationen zu speichern und zu verarbeiten: So sind z.B. Komplexe Deskriptoren sowohl für die TELDOK- als auch für die GOLEM-Anwendung verfügbar, während die Index-Generierung (Feinrecherche) ausschließlich bei der GOLEM-Anwendung möglich ist.

Indices werden bei der GOLEM-Anwendung auch genutzt, um sog. "nicht-terminale" syntaktische Strukturen und terminale morphosyntaktische Informationen zusätzlich zu syntaktischen Relationen aufzuzeichnen. Unter Ausnutzung dieser Indexinformation kann der Benutzer strukturelle Kontexte für die Deskriptoren auf der Ebene unterhalb des Dokuments spezifizieren. Diese Möglichkeit erlaubt dem Benutzer z.B., die kontextuelle Reichweite eines Ausdrucks "scharf einzustellen": Die Begriffe eines Ausdrucks stehen je nach der Reichweite in lockerer oder starker Assoziation.

Die Feinrecherche überbrückt somit die Lücke zwischen den bei den Hauptbereichen eines IR-Systems, d.h. zwischen dem Bereich der Dokumente und dem der Deskriptoren. So ist es prinzipiell möglich, auch bei diesem formal-inhaltlich orientierten Suchwortverfahren relativ umfangreiche Dokumente zu speichern, da die Recherche bei Bedarf über die Feinrecherche auf textuelle Untermengen von Dokumenten beschränkt werden kann. Übertragen auf die GOLEM-Anwendung würde dies die Identifizierung von Texteinheiten gestatten, die größer als der Satz sind (z.B. Paragraphen, Abschnitte, Kapitel), wofür die Feinrecherche-Funktion genutzt werden könnte.

In der vorliegenden CTX-Modellanwendung von GOLEM wird für die Feinrecherche die folgende Auswahl von Strukturmerkmalen (aus der ermittelten Text-/Satzstruktur) aufbewahrt:

| | |
|---|---|
| Satz | S |
| Subsatz | SB |
| Wortklasse | VRB,SUB,ADJ |
| GEN-Relation | GL, GR |
| PRP-Relation | PL, PR |
| ADJ-Relation | A |

Die linken und rechten Elemente genitivischer und präpositionaler Relationen beziehen sich auf die rechten und linken "Seiten" der entsprechenden GEN- und PRP-Relationen der Komplexen Deskriptoren. Bei der Adjektiv-Relation muss die Stellung nicht markiert werden, da die Textordnung (im Deutschen) festgelegt ist: Adjektiv-Attribute gehen den Modifikatoren immer voraus.

Theoretisch könnten nahezu die gesamten Ergebnisse der Sprachanalyse mit wenig Informationsverlust gespeichert werden. Das Problem liegt dann jedoch in der Formulierung von Suchbegriffen, besonders wenn komplexes Pattern-Matching gefordert wird. Beim Benutzer kann eher ein

Interesse an den begrifflichen Eigenschaften von Deskriptoren vorausgesetzt werden, während die syntaktischen Strukturen, in denen diese Deskriptoren vorkommen, ihm gleichgültig bzw. nicht für ihn "nachvollziehbar" sind; der GOLEM-Index jedoch erfordert eher einen expliziten Bezug auf die Oberflächen-Form des Textes als auf seine semantische Funktion.

Beispiel:
Text1: " .., der Schutz von Daten ... "
Text2: " ... Daten sind zu schützen ..."
Text3: " ... Daten werden vor Missbrauch geschützt ..."
Text4: " ... geschützte Daten missbräuchlich verwenden ..."

Die Information bleibt in diesen Paraphrasen im wesentlichen konstant, obwohl der syntaktische Kontext beträchtlich variiert. Der Benutzer (der den Text des Dokuments ja erst dann hat, wenn die Recherche erfolgreich war) kennt die formalen Charakteristika der Textoberfläche nicht und interessiert sich sicherlich auch nicht dafür. Stünde nur die Relation zwischen MISSBRAUCH und SCHUTZ zur Verfügung, erforderte jeder dieser Texte eine eigene Feinrecherche.

Für dieses Problem gibt es mehrere mögliche (Teil-)Lösungen. Erstens könnte das Analyseverfahren einschließlich der entsprechenden "Anreicherungsfunktionen" von CTX eine "tiefere" Darstellung liefern und dabei oberflächlich verschiedene Konstruktionen auf eine Art "kanonische Form" abbilden. Beispielsweise wird ja - wie erwähnt - gegenwärtig z.B. eine Passiv-Aktiv-Transformation vorgenommen, um das syntaktische "Tiefensubjekt" zu markieren.

Weiterhin könnte die maschinelle Analyse die Kasusgleichheit in nominalen und verbalen Gruppen erkennen, also (im Idealfall) z.B. die Identifizierung der Agens-Rolle gestatten. Schließlich könnten die nominalen Formen explizit zu "Tiefenstruktur"-Deskriptoren erklärt werden (wobei sie zusätzlich als substantivische Deskriptoren beibehalten würden).

Es darf jedoch nicht vergessen werden, dass das zugrunde liegende Analyseverfahren hierfür keine ausreichenden Bedingungen schafft. Es erschien auch wenig sinnvoll, zum gegenwärtigen Zeitpunkt CTX-eigene Vertiefungsverfahren zu entwickeln. (Man vgl. allerdings die Arbeit von THIEL 1982, die eine entsprechende Erweiterung des SUSY-Systems beschreibt). Hier bleiben letztlich die Ergebnisse computerlinguistischer Verfahren der 80-er Jahre abzuwarten, wie sie sich z.B. in den Planungen zu einem "Europäischen Übersetzungssystem" (EUROTRA) andeuten.

Eine Teillösung wäre es, die vorliegenden Ergebnisse der automatischen Sprachanalyse zu modifizieren. Man könnte z.B. der Syntaxstruktur Pseudo-Etiketten zuweisen, um potentielle Mitspieler-Rollen entsprechend kontextfreien Regeln darzustellen. So ist ein Subjekt in einem aktiven Satz oder ein Objekt nach der Präposition "von" in einem passivischen Satz meist ein Agens. Ähnliche Hypothesen sind für Mitspieler in Nominalisierungen möglich, die durch Derivationsrelationen im Thesaurus erkannt werden können (z.B. zwischen ANKUENDIGEN und ANKUENDIGUNG). Eine solche Möglichkeit wäre wesentlich unproblematischer als im MÜ-Verfahren selbst: Die Analyse beschäftigt sich mit der Bestimmung von tatsächlichen Etiketten, während dieses Verfahren sich nur mit der Zuweisung von potentiellen befassen würde. Ein erheblicher Vorteil dieser Lösung wäre, dass die notwendigen Erweiterungen sowohl vom IR-System als auch weitgehend vom Analysesystem unabhängig sind. Der Preis einer derartigen Lösung wäre eine beträchtliche Anzahl potentieller Etiketten, von denen manche im strengeren

Sinne "falsch" wären (sog. "Overkill" bzw. "Überindexierung"), die alle in die Deskriptorliste bzw. als indizierte Rollenindikatoren in die Indices aufgenommen werden müssten.

Eine dritte mögliche Teillösung liegt in der Erkenntnis, dass jede Recherche dieser Art im allgemeinen komplexe Ausdrücke erfordert. Die Formulierung dieser Ausdrücke könnte in hohem Maße maschinengestützt oder sogar vollautomatisch erfolgen. Wenn die Analyseergebnisse als oberflächensyntaktische Relationen und Klassen ausgedrückt werden, wie es augenblicklich bei dem zugrunde liegenden Verfahren der Fall ist, könnte ein zusätzliches Umsetzungsverfahren zwischen die Formulierung des Benutzers und den Input in ein Retrievalsystem (z.B. in GOLEM) geschaltet werden.

Der Baustein, der ein solches Umsetzungs- oder Generierungsverfahren bereitstellt, könnte entweder als Teil der Retrieval-Software in das jeweilige IR-System integriert oder als Teil des CTX-Systems realisiert werden. Eine Reihe von Gründen sprechen dagegen, auf der bestehenden IR-Software aufzubauen. Erstens bieten nicht alle IR-Systeme Schnittstellen, die es dem Systemanwender ermöglichen, eigene Funktionen in das System zu integrieren, so dass die Menge möglicher Fremdsysteme erheblich reduziert würde. Zweitens ist ein IR-eigener Code vermutlich stark von systeminternen Konventionen abhängig, was die Informationsdarstellung bei den Deskriptoren und dem Text betrifft; daher würde die Realisierung dieser CTX-Komponente in Methode und Effektivität von einem Fremdsystem zum anderen erheblich schwanken. Drittens würde die Retrievalstrategie von systemspezifischen Formalismen abhängen, und die Kontinuität von CTX zwischen verschiedenen IR-Systemen würde daher einem Benutzer uneinsichtig bleiben. Viertens schließlich ginge so die Gelegenheit verloren, die Verarbeitung dieser Informationen in andere Retrievalmethoden des CTX-Verfahrens zu integrieren.

Die zweite Alternative, d.h. die Realisierung einer Retrievalschnittstelle als Teil des CTX-Verfahrens mit Anpassungen an das jeweilige IR-Host-System, umgeht diese Nachteile und ist zusätzlich ein Baustein, der größtenteils von fremden IR-System unabhängig ist. Ein zusätzlicher Vorteil ist, dass diese Lösung die Forderung nach der Einheitlichkeit der Etiketten lösen könnte, wenn auch auf Kosten komplizierterer Ausdrücke in einer Feinrecherche (die aber für den Benutzer unsichtbar bleiben: Black-Box-Prinzip).

### I.2.8 Natürlichsprachige Retrieval-Schnittstelle (NATURA)

Je komplexer die Deskribierungsmöglichkeiten in einem formal-inhaltlichen Suchverfahren werden, desto schwieriger wird für den Benutzer - besonders für den gelegentlichen Benutzer - die Bewältigung der begrifflichen Komplexität bei der Recherche. Dies gilt aufgrund der Verfeinerung und Ausweitung der Möglichkeiten und Informationsdetails in besonders hohem Maße für das CTX-Verfahren, da der Benutzer diese verschiedenen Bausteine natürlich voll nutzen sollte. Um so wichtiger ist es, benutzerorientierte Verfahren bereitzustellen, die die Belastung beim Retrieval verringern.

Für das Modellsystem CTX wurde daher ein Verfahren vorgesehen, das dem Benutzer in der letztlich intendierten Ausbaustufe ermöglichen soll, Dokumente auf der Basis natürlichsprachiger Problembeschreibungen zu finden (daher das Kürzel "NATURA"). "Natürlichsprachig" bedeutet nicht eine irgendwie geartete natürlichsprachige Simulation formal festgelegter Ausdrücke im Sinne einer Programmiersprache wie COBOL. Der Input bei der "Suchanfrage" soll zudem nicht in Form von (natürlichsprachigen) Fragen (wie z.B. bei den Systemen der Künstlichen Intelli-

genz), sondern in Form von Aussage-Sätzen, Nominalphrasen oder Wortlisten (nicht Deskriptoren) vorliegen. Damit initiiert der Benutzer den Retrievalvorgang.

NATURA befindet sich noch auf einer einfachen Entwicklungsstufe. In dem Modellsystem liegt es vorläufig als eigenständiger Baustein vor, der die Aufgabe hat, Stichwörter aus einer Folge von (beliebigen) natürlichsprachigen Sätzen oder Ausdrücken analog der IR-Implementierung (in der realisierten Form auf dem GOLEM-Host) zu extrahieren. Unter Verwendung der gleichen Verfahrensstrategien wie bei der Indexierung der der Informationsbank zugrunde liegenden Dokumente werden die Problembeschreibungen auf Einfache und Komplexe Deskriptoren abgebildet. Einziger Unterschied ist, dass der CTX-Thesaurus bei der Bearbeitung eines dem System bislang unbekannten Stichworts nicht angepasst wird. Eine derartige Liste von Einfachen und Komplexen Stichwörtern bedeutet für den Benutzer schon eine erhebliche Erleichterung. So erfährt er auf einen Blick, in welcher Form welche Stichwörter als Deskriptoren in Frage kommen und in welcher äußeren Form sie für seine Recherche relevant sind: er muss z.B. nicht schon vorher "wissen", wie die Grundformen aussehen, welche mehrwortigen Begriffe von dem System als Feste Wendungen gespeichert werden und welche Bedeutungsvariante (bei verwendeten mehrdeutigen Wortformen) ggf. relevant ist. Transitive, kommutative und assoziative Erweiterungen aller KON-Relationen sind (zu Testzwecken) in der Bearbeitung der NATURA-Problembeschreibung eingeschlossen. Der Output von NATURA kann damit in diesem Entwicklungsstadium als eine Liste sachdienlicher Deskriptor-Vorschläge an den Benutzer zur Gestaltung seiner Suchanfrage-Strategie in dem IR-System betrachtet werden.

Als Erweiterung von NATURA könnte z.B. die Generierung von Deskriptoren-Indices zu den Einträgen in der Stichwortliste ins Auge gefasst werden. Diese könnten dem Benutzer dabei helfen, seine eigenen Feinrecherchefragen zu formulieren. Hier zeigt sich jedoch bereits, dass die Entwicklung einer natürlich-sprachigen Schnittstelle unabhängig von einem IR-Host nur bis zu einem gewissen Grade akzeptabel erscheint. Gegenwärtig bleibt der Output eine Sammlung taktischer Suchvorschläge, und der Benutzer ist für die Anwendung dieser Formulierungen im Format des IR-Formalismus selbst verantwortlich. Weiterhin werden letztlich alle Erweiterungen der augenblicklichen Anwendung durch die Eigenschaften des IR-Systems bestimmt, sobald ein derartiger Baustein in die spezifische IR-Software integriert wird.

Eine derartige Integration kann auf verschiedenen Stufen erfolgen. Die erste Stufe beinhaltet etwa die automatische Erstellung einer Liste von Deskriptoren, deren jeder einem Begriff in der Stichwortliste entspricht. Der Output dieser Funktion wäre im wesentlichen mit dem Output der jetzigen NATURA-Implementierung identisch. Bei einer Implementierung in der GOLEM-Software würde nur zusätzlich zu jedem Begriff ein numerischer Wert angezeigt, der die Anzahl der gespeicherten Dokumente angibt, die durch diesen Deskriptor identifiziert wurden. Der Benutzer könnte somit anhand der üblichen Formulierung von Retrieval-Kommandos entsprechend den Konventionen des IR-Systems weiter arbeiten: z.B. durch die Erweiterung der Deskriptorenliste über Thesaurusrelationen bereits aufgenommener Begriffe; durch Löschen von Begriffen aus der Deskriptorliste; durch die Formulierung logisch verknüpfter Ausdrücke, durch die Verwendung von Feinrecherche-Techniken mittels Rollenindikatoren; durch Ausgabe- und Unterdrückungsverfahren im Zusammenhang mit der Struktur der Datenbank und durch die Ausschöpfung aller übrigen Möglichkeiten des Systems.

Zur Zeit ist NATURA eher ein Mittel zur Demonstration einer bequemen Recherche und zu Testzwecken als im eigentlichen Sinne, ein "Baustein" zur Verwendung in einem IR-System.

Wie nützlich die automatische Erstellung einer Deskriptorenliste auch sein mag: Die ursprüngliche Aufgabe des Retrievals beruht immer noch auf den Formalismen und Denkprozessen, die von dem Fremdsystem gefordert werden.

Eine Beschreibung künftiger Entwicklungsstufen überschreitet den Rahmen dieser Untersuchung. Um jedoch zu verdeutlichen, in welche Richtung die weitere Entwicklung verlaufen wird, sollen zunächst eine Art "Idealziel" und danach einige der Schwierigkeiten beschrieben werden, die die Erreichung dieses Zieles alles andere als trivial erscheinen lassen - immer noch unter der Einschränkung gegenüber Systemen der "Künstlichen Intelligenz" (KI), dass es sich nach wie vor um ein formal-inhaltliches Stichwortverfahren handeln wird: also nur die Technik des Retrieval (Vermeidung von intellektueller Ermittlung des formalen Deskriptors, Reduktion der anderweitig erforderlichen, d.h. zu "lernenden" Kommandos) vereinfacht und benutzerfreundlicher gestaltet werden soll.

Im Idealfall sollte der Benutzer seine Problembeschreibung eingeben können und - als unmittelbare Antwort vom System - eine Reihe von Dokumenten erhalten, die in der Datenbank gefunden wurden (z.B. in der Reihenfolge der Relevanz, die über Relevanzfunktionen ähnlich zu CONDOR ermittelt werden). Die Ergebnisse könnten dem Benutzer sequentiell (am Bildschirm) angezeigt werden. Zusätzlich zu diesem Vorgehen auf "oberer" Ebene sollte der Benutzer jedoch jederzeit auf untere Ebenen zugreifen können, d.h. auf eher formale Suchverfahren, indem er seine eigenen (booleschen) Verknüpfungen auf der Grundlage der ermittelten potentiellen Deskriptoren eingibt. Auf der oberen Ebene der Recherche sollten jedoch vom Benutzer keine Kenntnisse von oder Verständnis für irgendwelche technischen Konzepte oder Abläufe erwartet werden, die mit dem Recherchevorgang zusammenhängen - einschließlich der Kenntnis der äußeren Gestalt von Deskriptoren, grammatischer Strukturen und Thesaurusrelationen. Auf dieser Ebene werden die textuellen Eingaben intern ausgewertet und dem Benutzer nur auf Anfrage ausgegeben.

Unter folgenden Bedingungen könnte zugleich auf den Thesaurus (stufenweise) vollautomatisch zugegriffen werden. Auf einer ersten Stufe würden die in der vom Benutzer eingegebenen Problembeschreibung verwendeten Wörter, die Thesauruseinträgen entsprechen, automatisch durch Deskriptoren ersetzt. Dabei ließen sich alle Synonyme, Langformen, Kurzformen, Akronyme, Abkürzungen, Schreibvarianten und ggf. Derivationen eines Begriffs in die Suche mit einbeziehen. Auf einer zweiten "Relevanzstufe" könnten Quasi-Synonyme, assoziierte Begriffe und Hyponyme mit entsprechend reduzierten Relevanzgewichtungen einbezogen werden. Kurz gesagt dient diese Art des Thesauruszugriffs dazu, die Vollständigkeit der Deskriptoren mit so wenig Abweichungen vom semantischen Gehalt des ursprünglichen Stichworts wie möglich zu verbessern. Die Verwendung des Thesaurus, z.B. zur Erweiterung oder Eingrenzung des semantischen Feldes, oder die Auflösung von Homonymen könnte über eine Interaktion (Präzisierungs- bzw. Verdeutlichungsdialog) zwischen Benutzer und System eingebracht werden.

Diese Interaktion ist auf zwei Ebenen vorstellbar, die man als "prozedural" und "konversationell" bezeichnen kann. Die prozedurale Interaktion bedeutet einen Rückgriff auf die Input-Phase der Problembeschreibung. Wenn der Benutzer mit den Retrievalergebnissen nicht zufrieden ist, sollte er entweder die Eingabe der Anfrage wiederholen (vielleicht aufgrund neuer Ideen beim Durchlesen der ausgegebenen Dokumente) oder aber die "alte" Anfrage verbessern können. Diese Verbesserung kann in Form der Markierung oder Akzentuierung einzelner Teile der Suchanfrage erfolgen, um die Berechnung der Relevanzwahrscheinlichkeit zu modifizieren. (Auch hierzu liegen

bei CONDOR einige interessante Überlegungen vor, vgl. CONDOR 1980).

Die konversationelle Interaktion würde dem Benutzer eine Kontrolle über die Bearbeitung seiner Suchanfrage erlauben und Entscheidungen ermöglichen, die das System selbst nicht tätigen kann. Unter folgenden Bedingungen könnte eine derartige Interaktion vorgesehen werden:

(1) Homographie/Homonymie

Unter bestimmten Umständen kann das System Mehrdeutigkeiten automatisch auflösen. Syntaktische Klassen z.B. können i.a. aus dem syntaktischen Kontext bestimmt werden, und in den Thesaurus eingebundene text-typologische Beschränkungen können mögliche Lesarten eines Homographen/Homonyms, die unterhalb einer gewissen Plausibilitätsschwelle liegen, ausschalten (s. Teil I.1).

Wo jedoch Homographen/Homonyme ungelöst verbleiben, kann das System automatisch ein Informationsdokument anzeigen, das mit dem entsprechenden Homographen/Homonym verknüpft ist, so dass der Benutzer die passende Lesart identifizieren kann. Varianten in diesem Verfahren existieren bereits in den Modellimplementierungen von CTX, wobei die standardmäßigen formalen Retrieval-Mechanismen der betreffenden IR-Systeme verwendet werden.

(2) Änderungen an semantischen Merkmalen

Unter Kontrolle des Benutzers könnte das System - analog zu ähnlichen Mechanismen der marktgängigen IR-Systeme - die semantischen Vernetzungen der Begriffe in der Suchanfrage erweitern, eingrenzen oder verschieben. "Auf Wunsch" könnten Stichwörter entsprechend den Thesaurus-Relationen aufgerufen werden, die noch nicht automatisch/systematisch in die Recherche einbezogen wurden, auch in Anlehnung an sekundäre Relationen, um z.B. die Synonyme eines automatisch ausgegebenen Quasi-Synonyms oder assoziierten Begriffs zu berücksichtigen. Wie bei der Homographie/Homonymie stellt diese Form der Interaktion lediglich die Modifikation eines Bausteins dar, der bereits zum Standardverfahren eines jeden IR-Systems gehört, das einen Thesaurus enthält. Der Unterschied liegt in dem Grad der Benutzerfreundlichkeit und natürlich darin, dass die Ergebnisse dieser Interaktion in den Retrieval-Mechanismus (Relevanzfunktion) des Systems integriert sind.

(3) Interaktionen, die durch einzelne Deskriptoren ausgelöst werden

Ist ein Dialog entweder nicht-produktiv oder über-produktiv, kann das System den Benutzer informieren und eine Verdeutlichung fordern. Als Hilfe für den Benutzer könnte das System zusätzliche Informationen ausgeben, die für eine Entscheidung sachdienlich wären. Wenn z.B. ein Deskriptor über-produktiv ist, könnte das System eine "Panne" in der Produktivität seiner Hyponyme melden; umgekehrt könnte ein unter-produktiver Deskriptor Hinweise zu allen assoziierten Begriffen, Quasi-Synonymen, Ganz- und Teilbegriffen geben, ob sie nun in der ursprünglichen Rechercheformulierung enthalten waren oder nicht.

(4) Interaktionen, die durch Kombinationen von Deskriptoren ausgelöst werden

Intellektuelle Suchstrategien beinhalten i.a. zunächst die Identifizierung großer Untergruppen, indem logische Verknüpfungen von Untergruppen verwendet werden, die durch Einzelde-

skriptoren repräsentiert sind. Darauf werden durch logische Disjunktionen selektierte Begriffe angewandt: Das Zwischenkriterium des "Erfolgs" ist die taktische Reduktion auf verwendbare (auswertbare) Größen. Diese Strategie wäre maschinell relativ einfach zu simulieren, aber ein Faktor fehlt: die intuitive Selektion der "besten" Kombination von Deskriptoren. Daher muss das System die Möglichkeit einer intellektuellen Überwachung und Regulation seiner Strategie der logischen Kombination von Deskriptoren zulassen.

(5) Die syntaktische Struktur der Suchanfrage

Das System kann nicht wissen, ob die gesamte Suchanfrage oder ob nur ein Teil davon für den Retrievalvorgang sachdienlich ist. Es kann auch nicht "zufällige" Ergebnisse der Tatsache, dass die Suchanfrage natürlichsprachig gestellt wurde, ausschließen. Der Benutzer muss daher die Möglichkeit haben, Informationen "zurückzuziehen", die er zunächst implizit über die natürlichsprachige Problemformulierung eingegeben hat. Dies kann durch Rückgriff auf die ursprüngliche Suchanfrage (d.h. durch "prozedurale" Interaktion) oder durch eine Unterbrechung der Recherche (durch "konversationelle" Interaktion) geschehen, wobei die Generierung von Feinrecherche-Ausdrücken unter Verwendung von Indices beeinflusst wird. Es ist zu erwarten, dass die künftige Entwicklung dieser Systemkomponente die Form eines Verfahrens annehmen wird, das sich anfangs stark auf die Hilfe des Benutzers stützt, dass aber die interaktiven Züge in dem Maße freier wählbar werden, in dem der Prozess der Formulierung der Suchanfrage verbessert wird.

Es bedarf noch erheblicher Grundlagenforschung, um nicht nur effektive automatische Suchstrategien zu entwickeln, sondern auch, um das Ausmaß zu erforschen, in dem syntaktische und semantische (relationale) Merkmale von Deskriptoren Auswirkungen auf die Weise haben, in der sie in den Suchstrategien verwendet werden.

Aus diesem Grund sollten andere Arten semantischer Information (z.B. die generative Semantik) zusätzlich zu den gewöhnlichen, thesaurusartigen Relationen in Betracht gezogen werden. Schließlich müssen Feldversuche ausgeführt werden, um die Auswirkungen "naiver" Benutzerfragen bzw. -reaktionen zu bestimmen sowie das Ausmaß, in dem der Benutzer lernen kann, die verschiedenen Interaktionsmöglichkeiten auszunutzen.

Harald H. Zimmermann, Edith Kroupa, Gerald Keil (1983):
C T X - Ein Verfahren zur computergestützten Texterschließung
TEIL II

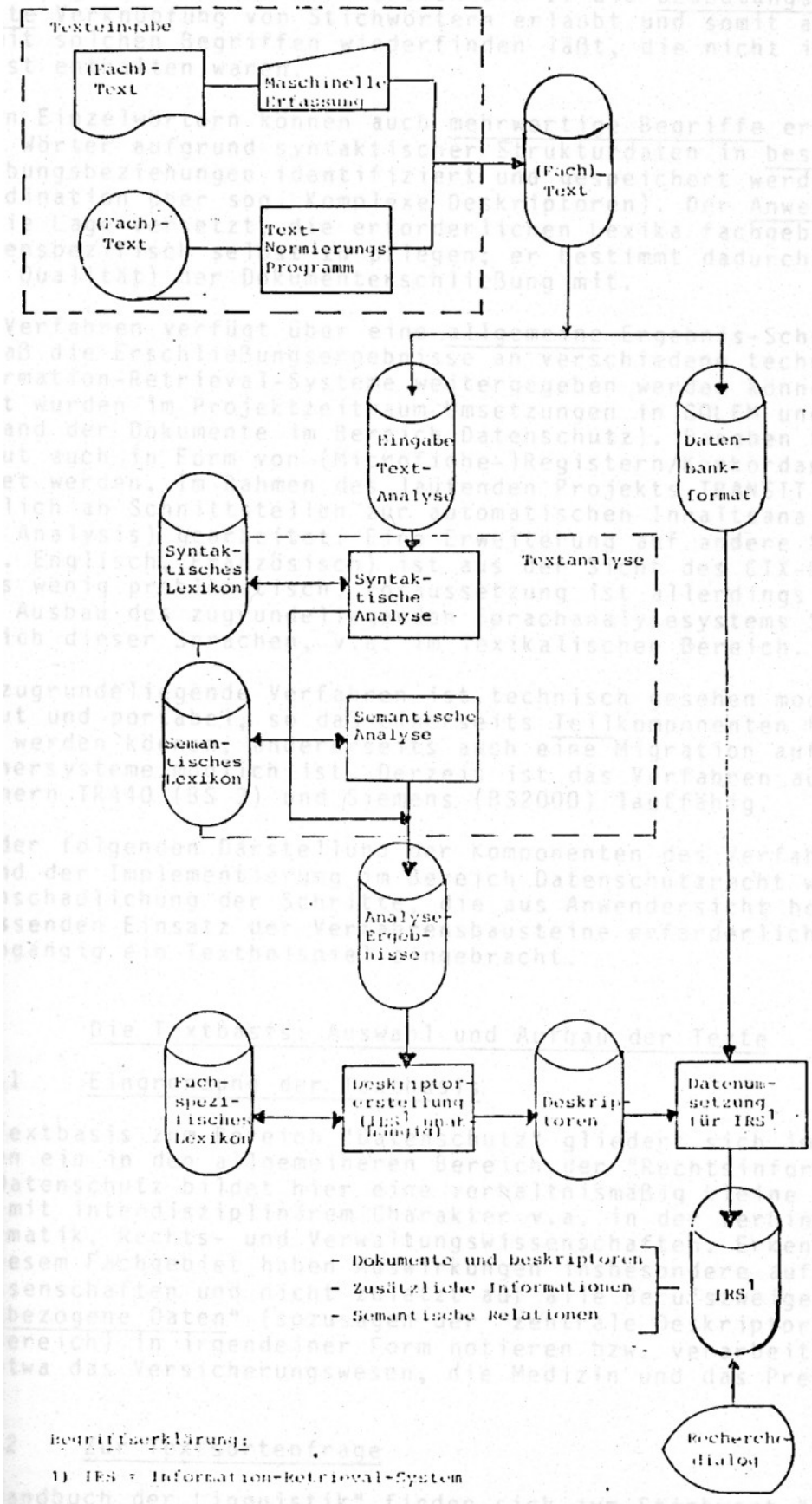
## II. LABORANWENDUNG VON CTX BEI RECHTSTEXTEN (Bereich Datenschutzrecht)

Während Teil I die allgemeine Problematik der computergestützten Textanalyse (I.1) und die wesentlichen Lösungsansätze der Forschungsgruppe (I.2) behandelt, soll im folgenden das in der 2. Projektphase realisierte Labormodell ausführlicher vorgestellt werden. Die Laboranwendung

bezog sich auf einen Teilbereich des Rechts ("Datenschutz"). Die Auswahl dieses Bereichs war eher zufällig. Im "Juristischen Informationssystem" (JURIS) ist dieser Bereich während der Projektlaufzeit nicht im Schwerpunkt bearbeitet worden (allerdings liegt das Bundesdatenschutzgesetz - BDSG - auch in einer GOLEM/PASSAT-Deskribierung vor). Umgekehrt konnte es von Vorteil sein, dass bei JURIS die allgemeine Problematik einer textorientierten maschinellen Rechtsdokumentation behandelt wurde. Im Projektverlauf hat es deswegen häufige Kontakte und Erfahrungsaustausch mit der JURIS-Entwicklungsgruppe und (auf den allgemeinen Benutzertreffen) auch mit JURIS-Benutzern gegeben.

Es war nicht beabsichtigt, die Laboranwendung für JURIS-Benutzer allgemein praktisch zugänglich zu machen: Bei gelegentlichen Demonstrationen konnten jedoch wertvolle Anregungen eingebracht werden. Projektziel war stets, den Bereich "Datenschutzrecht" als Exemplum zu sehen, der das Verfahren in seinen Grundlagen nicht festschreiben sollte. Inzwischen sind - nach Abschluss des Berichtszeitraums - in einem weiteren Forschungsprojekt "Transfer von Informationstechnologie" (TRANSIT) ähnliche "Vortests" des Systems in den Bereichen Patentwesen, allgemeine Bibliographien etc. in Arbeit. Über die dortige Verfahrensweise wird an anderer Stelle berichtet werden.

Ehe die Details der Laboranwendung vorgestellt werden, soll eine kurze Verfahrensübersicht die wesentlichen Funktionen des Systems verdeutlichen (vgl. Abb. II.1):




Begriffserklärung:

1) IRS = Information-Retrieval-System

Abb. II.1 Überblick über das Gesamtsystem

Unter dem Begriff CTX (Computergestützte Texterschließung) sind verschiedene Verfahrensbausteine zur Sprachdatenverarbeitung zusammengefasst, die sowohl am Sonderforschungsbereich "Elektronische Sprachforschung" (Teilprojekte "Maschinelle Übersetzung" und "Automatische Lemmatisierung") als auch an der Universität Regensburg und später an der Fachrichtung Informationswissenschaft der Universität des Saarlandes (Projekt "Juristische Dokumentanalyse" (JUDO) und Anwendungen im Bereich Datenschutz (JUDO-DS)) entwickelt wurden bzw. in Weiterentwicklung sind.

Im Sonderforschungsbereich 100 wurde unter anderem das Basissystem SUSY (Saarbrücker Übersetzungssystem) erstellt, das auch eine anwendungsnahe Teilkomponente, die "Saarbrücker Automatische Textanalyse", umfasst (vgl. Kap. I.2.3). Mit Hilfe der Verfahrensbausteine von CTX werden deutschsprachige Texte automatisch formal-inhaltlich erschlossen, indem die in den Texten enthaltenen sinntragenden Wörter auf ihre Grundform reduziert und ggf. in sinntragende Bestandteile zerlegt werden. Grundlage für die Sprachdatenverarbeitung sind einmal umfangreiche Lexika, zum anderen Verfahren, bei denen der (Satz-)Kontext eines Textwortes mit berücksichtigt wird. Das vorwiegend morphologisch-syntaktische Verfahren (CTX-I) kann um eine semantische, d.h. wortbedeutungsorientierte Analysekomponente erweitert werden (CTX-II). Anschließbar ist ein textorientiertes Thesaurusverfahren, das in der Version CTX-I die morphologische und in der Version CTX-II die bedeutungsorientierte Verknüpfung von Stichwörtern erlaubt und somit auch Texte mit solchen Begriffen wiederfinden lässt, die nicht im Text selbst enthalten waren.

Neben Einzelwörtern können auch mehrwortige Begriffe ermittelt bzw. Wörter aufgrund syntaktischer Strukturdaten in bestimmten Umgebungsbeziehungen identifiziert und gespeichert werden (Präkoordination über sog. Komplexe Deskriptoren). Der Anwender wird in die Lage versetzt, die erforderlichen Lexika fachgebiets- oder themenspezifisch selbst zu pflegen; er bestimmt dadurch die Tiefe (und Qualität) der Dokumenterschließung mit.

Das Verfahren verfügt über eine allgemeine Ergebnis-Schnittstelle, so dass die Erschließungsergebnisse an verschiedene technische Information-Retrieval-Systeme weitergegeben werden können. Realisiert wurden im Projektzeitraum Umsetzungen in GOLEM und TELDOK (anhand der Dokumente im Bereich Datenschutz). Daneben kann der Output auch in Form von (Microfiche-)Registern/Konkordanzen ausgewertet werden. Im Rahmen des laufenden Projekts TRANSIT wird zusätzlich an Schnittstellen zur automatischen Inhaltsanalyse (Content Analysis) gearbeitet. Eine Erweiterung auf andere Sprachen (z.B. Englisch, Französisch) ist aus der Sicht des CTX-Rahmensystems wenig problematisch, Voraussetzung ist allerdings der weitere Ausbau des zugrundeliegenden Sprachanalysesystems SUSY bezüglich dieser Sprachen, v.a. im lexikalischen Bereich.

Das zugrunde liegende Verfahren ist technisch gesehen modular aufgebaut und portabel, so dass einerseits Teilkomponenten herausgelöst werden können, andererseits auch eine Migration auf andere Rechnersysteme möglich ist. Derzeit ist das Verfahren auf den Rechnern TR440 (BS 3) und Siemens (BS2000) lauffähig.

Bei der folgenden Darstellung der Komponenten des Verfahrens CTX anhand der Implementie-

rung im Bereich Datenschutzrecht wird zur Veranschaulichung der Schritte, die aus Anwendersicht bei einem umfassenden Einsatz der Verfahrensbausteine erforderlich sind, durchgängig ein Textbeispiel eingebracht.

## II.1 Die Textbasis: Auswahl und Aufbau der Texte

### II.1.1 Eingrenzung der Textbasis

Die Textbasis zum Bereich "Datenschutz" gliedert sich im wesentlichen ein in den allgemeineren Bereich der "Rechtsinformatik". Der Datenschutz bildet hier eine verhältnismäßig kleine Teildisziplin mit interdisziplinärem Charakter v.a. in der Verbindung von Informatik, Rechts- und Verwaltungswissenschaften. Erkenntnisse in diesem Fachgebiet haben Auswirkungen insbesondere auf die Sozialwissenschaften und nicht zuletzt auf alle Berufszweige, die "personenbezogene Daten" (sozusagen der "zentrale Deskriptor" in diesem Bereich) in irgendeiner Form notieren bzw. verarbeiten müssen, wie etwa das Versicherungswesen, die Medizin und das Pressewesen.

### II.1.2 Zur Textsortenfrage

Im "Handbuch der Linguistik" finden sich zum Stichwort "Textsorte" folgende Angaben: Eine Textsorte ist eine "Teilmenge von Texten, die sich durch bestimmte relevante gemeinsame Merkmale beschreiben und von anderen Teilmengen von Texten abgrenzen lassen" (HANDBUCH 1975, S.496).

Die Textdaten zum Datenschutz wurden nach dieser allgemeinen Vorgabe in Textsorten eingeteilt, ohne dass dabei Anspruch auf Vollständigkeit und präzise Differenzierung erhoben werden kann. Die Gliederung nach Textsorten hat eher formalen Charakter.

Von der Textstruktur her wurden folgende Gruppierungen bzw. Untergruppierungen vorgenommen:

(1) aufbauend auf einer formalen expliziten Gliederung lassen sich unterscheiden:
(la) Gesetzentwürfe,
(1b) Rechtsvorschriften,
(1c) Judikate,
(1d) Anweisungen,
(le) Informationsbroschüren,
(1f) Tätigkeitsberichte;

(2) einer lexikalischen Ordnung folgend:
(2a) Lexika;

(3) relativ frei strukturiert, dennoch allgemeinen Konventionen folgend, werden unterschieden:
(3a) Zeitungs- und Zeitschriftenartikel,
(3b) Fachaufsätze,
(3c) Fachbücher,
(3d) Rezensionen,
(3e) Abstracts,

(3f) Berichte (z.B. des BMFT, der KGST),
(3g) Informationsblätter;

(4) eine Mischform aus (2), gemäß einer lexikalischen Ordnung bei zahlreichen Erläuterungen, und (3), d.h. frei strukturiert bei einzelnen herausgegriffenen Erläuterungen, sind:
(4a) Technische Beschreibungen,
(4b) Technische Normen,
(4c) Kommentare.

Aus dem Bereich Datenschutz wurden v.a. als Material für zu verarbeitende Daten herangezogen:

- Gesetzentwürfe

Erfasst wurden Gesetzentwürfe, die von verschiedenen Parteien und Fraktionen, Ministerien, Vereinen, Vereinigungen, Verbänden oder Gewerkschaften als Vorschläge eingebracht wurden.
Zu unterscheiden wäre ggf. noch zwischen Gesetzentwürfen, die vom Parlament debattiert werden, und anderen, die Vorschläge für ein neues Gesetz enthalten (Anzahl in der Bundesrepublik jährlich zwischen 5 und 15).

- Rechtsvorschriften

Erfasst wurden das Bundesdatenschutzgesetz (BDSG) und einige Landesdatenschutzgesetze. Rechtsvorschriften für spezielle Bereiche (z.B. Gesundheitswesen, Geheimbereich, Meldewesen) stehen noch aus. Nur explizit als Datenschutzvorschrift benannte Vorschriften werden also berücksichtigt (jährlich ist hierzu mit bis zu 20 Gesetzen, Gesetzesänderungen und Verordnungen zu rechnen).

- Judikate

In diesen Bereich fallen juristische Entscheidungen, die im weiteren Sinn Datenverarbeitung und/oder Datenschutz behandeln (jährlich 10 bis 20).

- Kommentare

In Abhängigkeit von neuen Rechtsvorschriften erscheinen eine gewisse Zeit, verstärkt nach der Verabschiedung eines Gesetzes, kommentierte Fassungen (bis zu 20 pro Gesetz, pro Jahr ca. 5 Kommentare).

- Stenographische Berichte
Derartige Berichte werden im Anschluss an kleine und große Anfragen sowie Parlamentsdebatten anderer Art und Abstimmungen im Parlament erstellt (jährlich bis zu 40 im Bereich Datenschutz).

- Technische Beschreibungen, technische Normen, Tabellen. Sie erscheinen hauptsächlich in Monographien (jährlich ca. 5).

- Anweisungen

Dienstanweisungen erfolgen normalerweise verwaltungsintern und werden sehr selten veröffentlicht.

- Berichte (z.B.des BMFT, der KGST)

- Fachbücher

Jährlich erscheinen gegenwärtig ca. 80 deutschsprachige Monographien, die zum Bereich Datenschutz gerechnet werden können.

- Buchbesprechungen

Rezensionen erscheinen hauptsächlich in Fachzeitschriften und Fachzeitungen (jährlich sind 20 bis 100 Rezensionen zu erwarten).

- Fachaufsätze

Zieht man nur die einschlägigen (ca. 6 bis 12) Zeitschriften heran, so wird man etwa 250 Quellen erhalten; versucht man hingegen umfassend zu sein (etwa unter Integration der in der KJB verzeichneten 63 Fachzeitschriften für den Bereich Rechtsinformatik), so können ca. 600 (auch fremdsprachige) Artikel registriert werden (jährlich sind bis zu 400 deutschsprachige Fachartikel zu erwarten).

- Informationsblätter und -broschüren

Die Datenschutzbeauftragten des Bundes und der Länder sowie etwa Gewerkschaften, Vereine, Verbände usf. gehen zur Information der Bürger (der "Betroffenen") Broschüren heraus (jährlich bis zu 15).

- Zeitungsartikel und populärwissenschaftl. Literatur (jährlich ca. 200) .

Für eine umfassende Dokumentation zum Datenschutz von Relevanz sind zusätzlich die Abstracts, die z.B. eine Übersicht über ein Fachbuch geben, und Inhaltsverzeichnisse; in die Laboranwendung eingegangen ist zusätzlich die Textsorte "Monographie" (in Auszugsform)

### II.1.3 Textauswahl

Im 1. Projektzeitraum (vom Juli 1977 bis Dezember 1979) wurden zunächst

5 Gesetzesentwürfe
3 Normen sowie
2 Regelungen (Dienstanweisungen)

maschinell verarbeitet und testweise in eine Datenbank eingespielt.

Für die Testimplementierung in der 2. Projektphase (Januar 1980 bis Februar 1982) wurden die erfassten Texte in Gruppen eingeteilt, die im Projektzeitraum entsprechend für 2 Testphasen vor-

bereitet wurden (vorgesehen war noch eine 3. Phase, die jedoch aus projektökonomischen Gründen im Hinblick auf zu erwartende Ergebnisse im Projekt TRANSIT zurückgestellt werden konnte):

Texte der Phase 1

Diese Textmenge baut weitgehend auf den Daten zum 1. Projektzeitraum auf, reduziert jedoch um einige parallele Entwurfstexte. Die Texte waren zunächst - z.T. noch unter Verwendung wenig ausgereifter Algorithmen - analysiert und in eine Probe-Datenbank (TELDOK) eingebracht worden. Sie wurden im 2. Projektzeitraum erneut bearbeitet, um zusammen mit den Texten der Phase 2 eine homogene Datenbank (GOLEM) zu erhalten.

| | Anzahl Texte | Anzahl Dokum. | Anzahl Zeilen | Anzahl Token | Anzahl Sätze | durchschn. Satzlänge |
|---|---|---|---|---|---|---|
| Entwürfe | 2 | 40 | 884 | 3204 | 234 | 13,69 |
| Normen | 3 | 97 | 2518 | 10132 | 536 | 18,90 |
| Regelungen | 2 | 19 | 373 | 1447 | 75 | 19,29 |
| insgesamt: | 7 | 114 | 3775 | 14783 | 845 | - |

Texte der Phase 2:

Diese Textmenge bildet die eigentliche Grundlage zur Ermittlung der wesentlichen Eckdaten für eine erste Bewertung der Leistungsmöglichkeiten des CTX-Systems.

| | Anzahl Texte | Anzahl Dokum. | Anzahl Zeilen | Anzahl Token | Anzahl Sätze | durchschn. Satzlänge |
|---|---|---|---|---|---|---|
| Normen | 10 | 156 | 2894 | 27153 | 1192 | 22,70 |
| Judikate | 17 | 17 | 1475 | 8420 | 352 | 23,92 |
| Komm.auszug | 1 | 1 | 12 | 330 | 10 | 33,00 |
| Regelungen | 3 | 14 | 338 | 2886 | 126 | 22,90 |
| Tätigk.ber. | 1 | 183 | 6118 | 47419 | 2250 | 21,07 |
| Lexika | 1 | 551 | 5274 | 29997 | 2034 | 14,74 |
| Inf.brosch. | 1 | 85 | 2534 | 15097 | 902 | 16,73 |
| Fachaufs. | 1 | 1 | 70 | 623 | 31 | 20,09 |
| Zeitungsart. | 5 | 4 | 210 | 1429 | 77 | 18,55 |
| insgesamt: | 38 | 1012 | 18925 | 133354 | 6974 | - |

1) Token = laufende Wortformen

II.1.4 Erfassungskonventionen

Im folgenden wird das Textbeispiel zunächst im Original (Fotokopie) vorgestellt:

9.2 Datenschutzprobleme

Der Datenschutz ist nur ein - allerdings bedeutsamer - Aspekt der zahlreichen Probleme, die im Zusammenhang mit den "Neuen Medien" bedacht werden müssen. Feststellen läßt sich bereits jetzt, daß die rechnerunterstützten Telekommunikationsverfahren zu umfangreichen Sammlungen personenbezogener Daten in den Betriebszentralen (Bildschirmtext- oder Kabelzentrale) und bei privaten oder öffentlichen Anbietern führen können. Die selektive Auswahl aus einem großen Text- oder Bildangebot sowie die Kommunikation zwischen Benutzern und Programmanbietern ermöglicht Datenprofile, aus denen auf Sachverhalte aus dem Privatbereich des einzelnen geschlossen werden kann: seine Anwesenheit zu Hause, seine bevorzugten Programme, Reise-, Lektüre- oder sonstige Konsumgewohnheiten, Geldgeschäfte, Absender und Empfänger von Btx-Briefen, Hobbys, Lernverhalten, Geschicklichkeit im Umgang mit den Medien. Durch Zusammenfassung der Daten können umfassende Persönlichkeits- und Interessenprofile entstehen. Die Daten könnten auch für viele andere Zwecke zur Verfügung stehen; sie sind von größtem Interesse für Privatwirtschaft, öffentliche Verwaltung und Politik. Das Gefährdungspotential muß zwangsläufig wachsen, je mehr sich der technisch absehbare Trend zur individuellen, interaktiven Beteiligung des einzelnen verwirklicht. Die technische Möglichkeit der Kommunikationsfreiheit, die bereits als individuelles Grundrecht auf der Basis von Einzelaspekten der Meinungsfreiheit postuliert wird, bedarf der datenschutzrechtlichen Absicherung. Denn die Aktivität des einzelnen mit Unterstützung der Kommunikationstechnik hinterläßt Spuren, die nicht etwa nur die Überwachung, sondern auch seine Manipulation ermöglichen.

Für die Feldversuche des Dienstes "Bildschirmtext" sind in Berlin und Nordrhein-Westfalen spezialgesetzliche Landesregelungen getroffen worden. Im Interesse der schutzwürdigen Belange des Bürgers sind solche besonderen Rechtsvorschriften bei der endgültigen Einführung "Neuer Medien" erst recht erforderlich. Um eine bundeseinheitliche landesgesetzliche Regelung zu gewährleisten, ist für "Bildschirmtext" beabsichtigt, einen Länder-Staatsvertrag abzuschließen. Die Datenschutzbeauftragten des Bundes und der Länder haben hierzu einen Ergänzungsvorschlag zu den Datenschutzproblemen erarbeitet (Anlage 4 ).

Abb. II.2: Textbeispiel in Fotokopie

Die Texte wurden zunächst maschinenlesbar erfasst und anschließend so aufbereitet, dass sie dem internen Format des Analyseverfahrens entsprachen. Ein solcher Text kann gegenwärtig mit Groß-/Kleinbeschreibung, mit oder ohne Auflösung der Umlaute (ae oder ä, etc.) erfasst werden.

Besondere (wichtige) Markierungen (Sonderzeichen) sind:

* = Satzendepunkt: Zur Unterscheidung zum Punkt bei Abkürzungen und dem Dezimalpunkt muss der Satzendepunkt eindeutig markiert werden.

[ = "künstliches" Satzende: In allen Fällen, in denen im Text ein logisches Satzende ohne entsprechende Markierung vorliegt (dies gilt in erster Linie für das Ende von Überschriften), muss dieses Zeichen gesetzt werden.

Daneben sind besondere Regelungen getroffen für die Verarbeitung von Spiegelstrichen (Aufzählungen) und von Begriffen, die in Klammern stehen. Im einzelnen gibt hierzu das Handbuch zur Saarbrücker automatischen Textanalyse Auskunft (SATAN-Handbuch, Kap.-0).

Markierungen für die Identifizierung eines Textes und die Einteilung in recherchierbare Einheiten sind:

| | |
|---|---|
| $TXT /Kürzel/ | Beginn eines Textes; /Kürzel/ = Kennzeichen für den Text (z.B. Aktennummer) |
| $DOK | Beginn eines Textabschnitts (recherchierbare Einheit) |

So wurde das Textbeispiel wie folgt erfasst:

```
$TXTBEISPIEL
$DOK
9.2 Datenschutzprobleme[
Der Datenschutz ist nur ein - allerdings bedeutsamer - Aspekt der
zahlreichen Probleme, die im Zusammenhang mit den 'Neuen Medien' bedacht
werden muessen* Feststellen laesst sich bereits jetzt, dass die
rechnerunterstuetzten Telekommunikationsverfahren zu umfangreichen
Sammlungen personenbezogener Daten in den Betriebszentralen
(Bildschirmtextzentrale oder Kabelzentrale) und bei privaten oder
oeffentlichen Anbietern fuehren koennen* Die selektive Auswahl aus einem
grossen Textangebot oder Bildangebot sowie die Kommunikation zwischen
Benutzern und Programmanbietern ermoeglicht Datenprofile, aus denen auf
Sachverhalte aus dem Privatbereich des einzelnen geschlossen werden
kann: seine Anwesenheit zu Haus, seine bevorzugten Programme,
Reisegewohnheiten, Lektueregewohnheiten oder sonstige
Konsumgewohnheiten, Geldgeschaefte, Absender und Empfaenger von BTX-
Briefen, Hobbys, Lernverhalten, Geschicklichkeit im Umgang mit den
Medien* Durch Zusammenfassung der Daten koennen umfassende
Persoenlichkeitsprofile und Interessenprofile entstehen* Die Daten
koennten auch fuer viele andere Zwecke zur Verfuegung stehen; sie sind
von groesstem Interesse fuer Privatwirtschaft, oeffentliche Verwaltung
und Politik* Das Gefaehrdungspotential muss zwangslaeufig wachsen, je
mehr sich der technisch absehbare Trend zur individuellen interaktiven
```

```
Beteiligung des einzelnen verwirklicht* Die technische Moeglichkeit der
Kommunikationsfreiheit, die bereits als individuelles Grundrecht auf der
Basis von  Einzelaspekten der Meinungsfreiheit postuliert wird, bedarf
der datenschutzrechtlichen Absicherung* Denn die Aktivitaet des
einzelnen mit Unterstuetzung der Kommunikationstechnik hinterlaesst
Spuren, die nicht etwa nur die Ueberwachung, sondern auch seine
Manipulation ermoeglichen*
$DOK
Fuer die Feldversuche des Dienstes 'Bildschirmtext' sind in Berlin und
Nordrhein-Westfalen spezialgesetzliche Landesregelungen, getroffen
worden* Im Interesse der schutzwuerdigen Belange des Buergers sind
solche besonderen Rechtsvorschriften bei der endgueltigen Einfuehrung
'Neuer Medien' erst recht erforderlich* Um eine bundeseinheitliche
landesgesetzliche Regelung zu gewaehrleisten, ist fuer 'Bildschirmtext'
beabsichtigt,  einen  Laender-Staatsvertrag  abzuschliessen*  Die
Datenschutzbeauftragten des Bundes und der Laender haben hierzu einen
Ergaenzungsvorschlag zu den Datenschutzproblemen erarbeitet*
```

Abb. II.3: Erfassungsformat des Textbeispiels

## II.2 Textbezogene Wörterbucharbeiten in der Laboranwendung

### II.2.1 Die Wörterbücher im System CTX

Die Verfahrensweise von CTX ist in zweifacher Hinsicht lexikonorientiert: Wörterbücher werden eingesetzt zur Ermittlung von formal-inhaltlichen Stichwörtern bei der Texterschließung und - v.a. ergänzt um semantische Begriffsrelationen - zur Unterstützung der Textwiedergewinnung. Der Stichwortextraktion liegt - wie erwähnt - die deutsche Analysekomponente des Saarbrücker automatischen Übersetzungssystems SUSY zugrunde. Dieses System basiert auf zwei Wörterbüchern, einem weitgehend allgemeinsprachlichen morpho-syntaktischen Wörterbuch und einem (z.T. fachgebietsbezogenen) semantischen Lexikon:

Das morphosyntaktische Wörterbuch liefert Informationen zur maschinellen Zuordnung von Flexionsformen zu möglichen Grundformen und zum syntaktischen Verhalten der Wörter. Das semantische Wörterbuch enthält Kodierungen und Regeln, die eine automatische Auflösung von semantischen Mehrdeutigkeiten sowie die Zusammenführung von im Satzkontext (auch diskontinuierlich) auftretenden mehrwortigen Begriffen ermöglichen sollen (zu den Einzelheiten vgl. Kap. 11.3.3).

In Zusammenarbeit mit dem Sonderforschungsbereich "Elektronische Sprachforschung", Teilprojekt A2, das für die Grundlagenentwicklung des MÜ-Systems verantwortlich ist, wurden diese beiden Wörterbücher gepflegt: Wenn nötig, wurden sie an die Erfordernisse der Analyse der Testtexte durch Lexikonerweiterungen bzw. -korrekturen angepasst.

In Ergänzung zu diesen beiden Wörterbüchern wurde als eigenständige Entwicklung im Projektverlauf für die Texterschließung ein dritter Wörterbuchbereich mit fachgebietsrelevanten Informationen konzipiert und modellhaft realisiert. Integriert in dieses fachspezifische Lexikon ist der sog. "CTX-Thesaurus", ein System zur Darstellung und Aufbereitung begrifflicher Relationen. Im Labormodell orientierte sich dieser Thesaurus vorwiegend an dem Fachwortschatz zum Datenschutz.

In einem Teilsystem des Fachlexikons werden zunächst semantisch mehrdeutige Wörter gesammelt und entsprechend ihren Bedeutungsvarianten paraphrasiert. Die Einträge dieses Lexikons enthalten fachsprachliche Markierungen zu den durch das Sprachanalyseverfahren gewonnenen (ggf. vereindeutigten) Stichwörtern, der CTX-Thesaurus beschreibt morphosyntaktische Beziehungen (Derivationen) sowie semantische Relationen des deskriptorrelevanten Wortmaterials.

Während die beiden Analysewörterbücher im System CTX der allgemeinen linguistischen Erschließung der Texte, d.h. der Deskriptorerstellung, dienen und somit ohne größere Probleme auch in anderen Fachgebieten verwendet werden können, müssen das Fachlexikon und der darin integrierte CTX-Thesaurus naturgemäß jeweils fachgebietsspezifisch angepasst werden. Diese Daten finden zudem zweifach Anwendung: Ein Teil der im Fachlexikon gesammelten Informationen trägt bei zum Aufbau der formal-inhaltlichen Deskriptoren (vgl. Kap. II.4.2.4); darüber hinaus leisten die Begriffsrelationen dem Benutzer Hilfestellung beim Retrieval.

Alle Wörterbücher werden intellektuell aufgebaut und gepflegt, wobei allerdings so weit wie möglich computergestützte Verfahren zum Einsatz kommen. Die Einträge bauen in keinem Falle auf rein statistischen Auswertungen von Texten auf. Dieses Vorgehen berücksichtigt somit, dass durch intellektuelle Kodierung bzw. Kontrolle die Qualität der Dokumentbeschreibung verbessert werden kann.

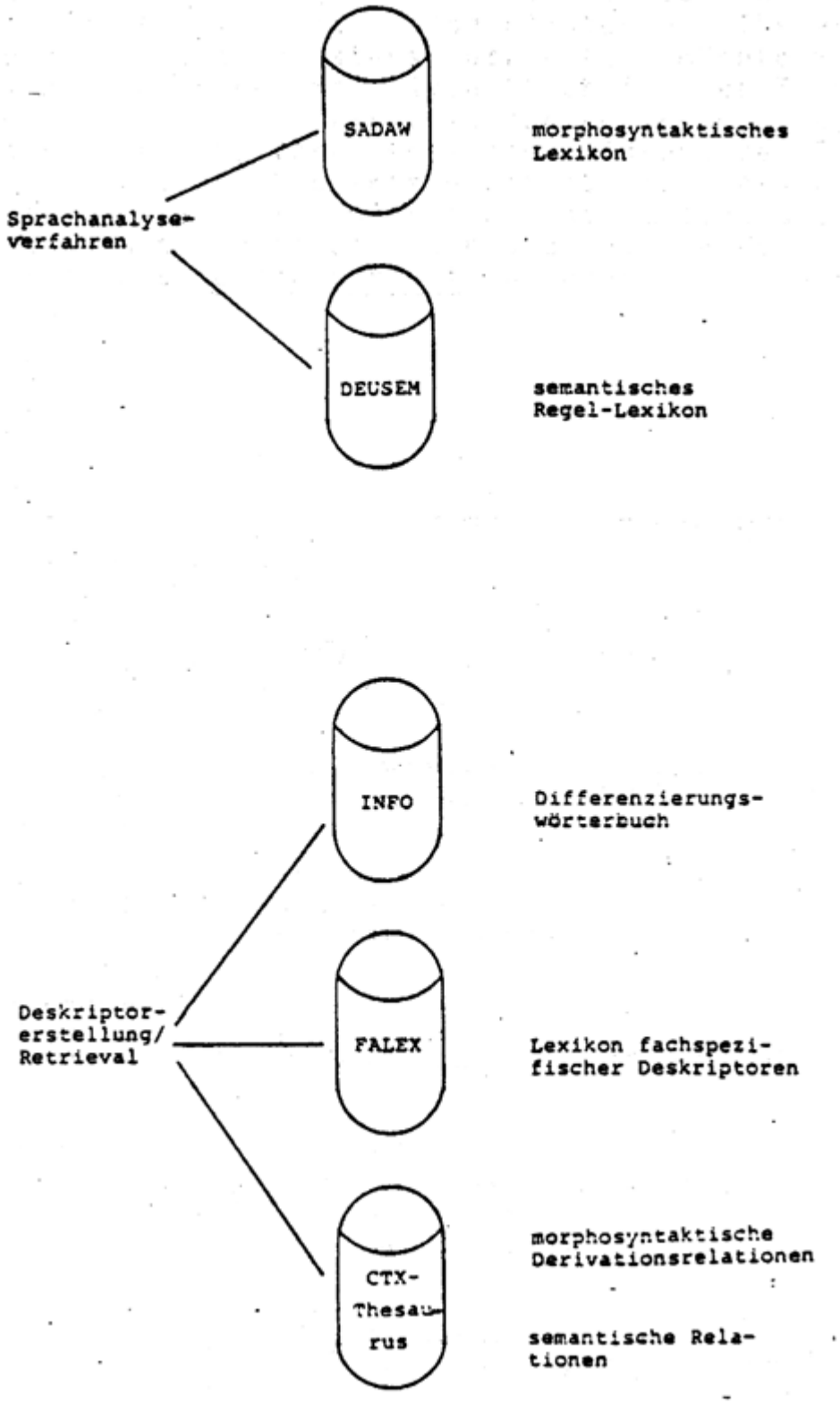



Abb. II.4: Übersicht über die Wörterbücher im System CTX

### II.2.2 Textabhängige Lexikonaufbereitung

Bei Aufbau und Pflege von intellektuell erstellten Wörterbüchern als Hilfsmittel zur automatischen Indexierung sind - falls nicht auf vorhandenes Material zurückgegriffen werden kann - im wesentlichen zwei Wege denkbar. Eine systematische, über das aktuell zu verarbeitende Textmaterial hinausgehende Wörterbucherweiterung kann den später anfallenden Aufwand weitgehend vorwegnehmen und ist sowohl der Vollständigkeit als auch der Konsistenz der im Einsatz befindlichen Lexika förderlich. Im allgemeinen liegt - auch fachgebietsspezifisch - genügend aufbereitetes "herkömmliches" Grundmaterial in Form von umfangreichen allgemeinsprachlichen und fachsprachlichen Lexika, von Terminologielisten und Thesauri vor (wenn auch letzteres im Bereich Datenschutz - fast möchte man sagen: ausnahmsweise - wegen der relativen "Neuheit" des Themenbereichs nicht in dem Maße der Fall war). In der konkreten Projektsituation war diese systematische Vorgehensweise aufgrund der Finanz- und Personalausstattung des Projekts nicht möglich, das Forschungsprojekt war zudem auf eine exemplarische Entwicklung und -anwendung ausgerichtet. Somit wurden die fachsprachlichen Wörterbücher im Rahmen des Projekts textorientiert aufgebaut und ergänzt.

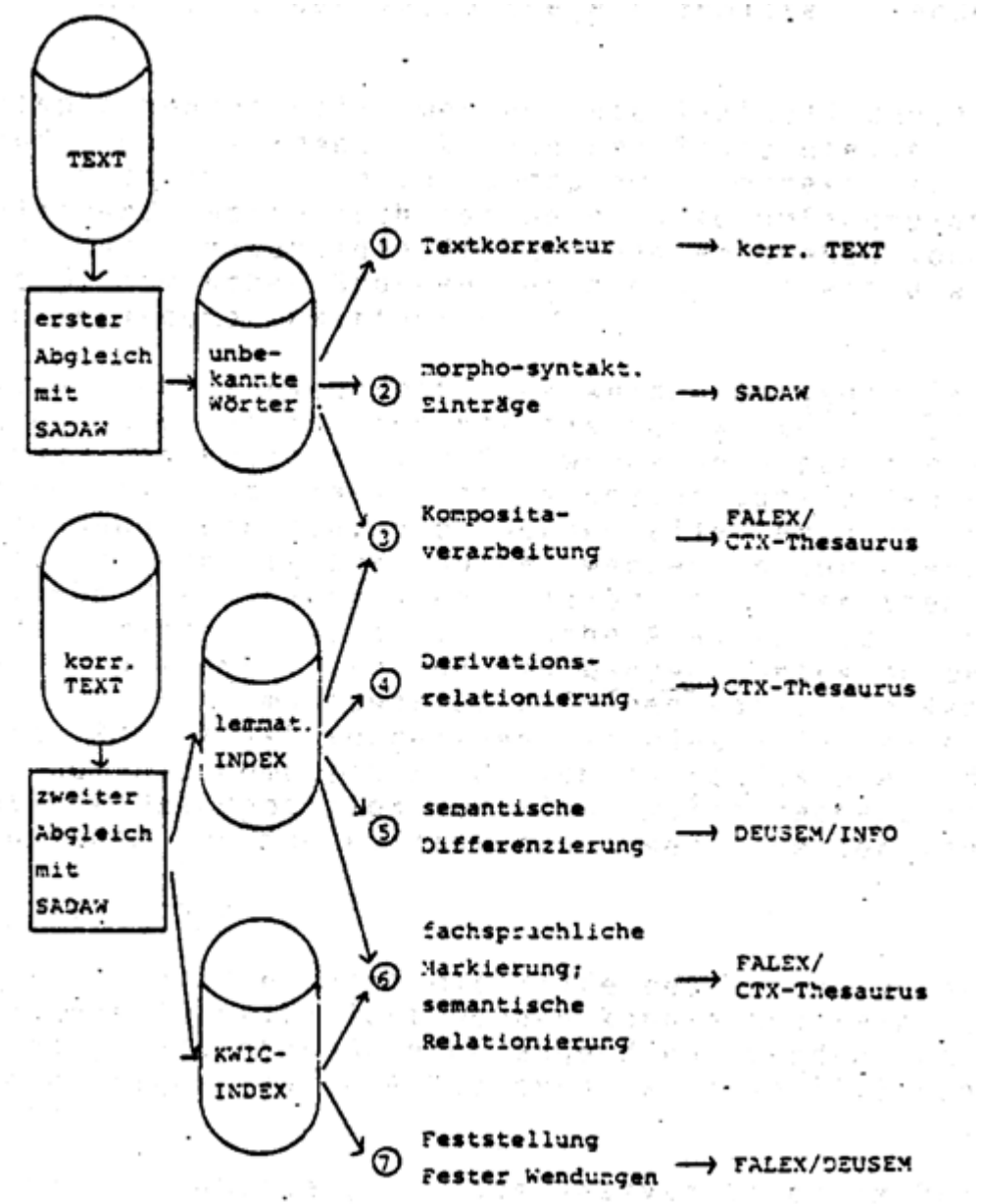



Abb. II.5: Übersicht über die Wörterbucharbeiten im System CTX

Im folgenden werden die zur Erschließung eines neuen Textes anfallenden Wörterbucharbeiten anhand eines Beispieltextes erläutert.

II.2.2.1 Erweiterung des morphosyntaktischen Wörterbuchs

Der Grad der Vollständigkeit des morphosyntaktischen Wörterbuchs wirkt sich besonders stark auf die Qualität der Analyseergebnisse aus. In einem Abgleich des Textes mit diesem Wörterbuch werden daher zunächst alle Textwortformen ermittelt, deren Grundformen nicht bereits im Lexikon enthalten sind. Die entsprechenden Textwortformen werden dabei in einer speziellen Liste festgehalten. Die Liste kann umfassen:

(1) Erfassungsfehler (Rechtschreibfehler)

Die Erfassungsfehler werden anschließend im Text korrigiert. Nicht erkannt werden auf diese Weise natürlich Erfassungsfehler wie KARTEI anstelle von DATEI.

(2) Wortzusammensetzungen

Hierbei handelt es sich um Wörter, die selbst nicht als Grundformen/Stammformen im Wörterbuch enthalten sind, deren Bestandteile jedoch alle enthalten sind oder über Dekompositions- bzw. Derivationsregeln ermittelt werden konnten.

Die entsprechenden Fälle werden intellektuell daraufhin überprüft, ob die Identifikation der Bestandteile (d.h. Stämme/Grundformen/Ableitungsmorpheme) korrekt ist. Das zugrundeliegende Wörterbuch ist bereits so umfangreich, dass durchaus "unsinnige" Identifikationen entstehen können, zumal aus ökonomischen Gründen nur die systemseitig wahrscheinlichste Zerlegung ermittelt wird.

Offensichtliche Zerlegungsfehler werden in jedem Falle durch eine Ergänzung des morphosyntaktischen Wörterbuchs bereinigt. (Im Rahmen des Projekts TRANSIT wurde inzwischen diesem Problem durch die Integration eines Kompositum-Zerlegungs-Lexikons Rechnung getragen. Hier werden alle Zerlegungen aufbewahrt, fehlerhafte Zerlegungen werden entsprechend korrigiert). Für die übrigen identifizierten Wörter steht dies frei, da zumindest die syntaktische Analyse i.d.R. unbeschadet der Nicht-Identifikation des ganzen Wortes aufgrund der Angaben des Bestimmungswortes erfolgen kann. (Im Rahmen des Projekts TRANSIT ist vorgesehen, in diesen Fällen einen Zugriff auf das Kompositum-Lexikon zu realisieren, da die automatische Dekomposition verhältnismäßig rechenzeitintensiv ist und zudem der Kontrollaufwand gesenkt werden kann.)

(3) Unbekannte Wörter

Textwortformen, die nicht mit Hilfe des morphosyntaktischen Wörterbuches identifiziert werden können und keiner der beiden oben genannten Gruppen angehören, werden kodiert und nachgetragen (auf die Kodierung wird in II.3 näher eingegangen).

Es ist hier anzumerken, dass das Analysesystem im Prinzip auch Sätze syntaktisch weiterverarbeiten kann, die "unbekannte" Wörter enthalten, da in jedem Falle aufgrund der Zeichenstruktur und ggf. sonst vorhandener Angaben wie Groß-/Kleinschreibung syntaktische Informationen vergeben werden. Allerdings ist eine korrekte Grundformenermittlung oder gar eine semantische Vereindeutigung dabei nicht gewährleistet. Somit ist CTX im Grunde auch darauf ausgerichtet, Texte wartungsfrei (bezüglich der Wörterbuchpflege) zu indexieren. Das setzt natürlich eine recht "gesättigte" Lexikonbasis voraus, bei der der Anteil "unbekannter" Wörter unter einer bestimmten Akzeptanzschwelle liegt. Umgekehrt sinkt naturgemäß der entsprechende Pflegeaufwand, so dass eine Lexikonpflege mit wachsendem Inventar kostenmäßig kaum mehr ins Gewicht fallen wird.

Für die Deskriptorzuteilung werden nur Textwortformen der Wortklassen Substantiv, Verb und Adjektiv (der so genannten Hauptwortklassen) berücksichtigt. Da die Ergebnisse der Analyse bei dem vorliegenden Verfahren weitreichende Auswirkungen haben, z.B. auf die semantische Vereindeutigung und die Erstellung von Komplexen Deskriptoren, (vgl. Kap. II.4.3), die erreichbare

Qualität aber wesentlich von der Konsistenz des syntaktischen Lexikons abhängig ist, wurde das morphosyntaktische Lexikon in der Laboranwendung "Datenschutzrecht" um alle Einträge der Gruppe (3) und nach erfolgter Kontrolle der ermittelten Wortzusammensetzungen um fehlerhaft zerlegte Wörter aus Gruppe (2) erweitert.

Beispiele für Rechtschreibfehler, die mithilfe des Lexikons festgestellt worden sind:

AUSGEWAHLTEN
ANMELDUNGNACH
UEBERLICK

Beispiele für die automatische Identifikation zusammengesetzter Wörter:

| | |
|---|---|
| BESTELLFORMULAREN | BESTELL/FORMULAR |
| VERFAHRENSENTWUERFEN | VERFAHREN*S/ENTWURF |

Beispiele für Wörter der Gruppe (3), die dem System unbekannt waren, insoweit der Stamm oder ein Bestandteil (zuvor) noch nicht im Lexikon waren:

| | |
|---|---|
| DIREKTUPDATING | (Updating nicht im Lexikon) |
| INFORMATIK | (Informatik " " ) |
| KEINERLEI | (keinerlei " " ) |

Im Textbeispiel wurden folgende Wörter entsprechend identifiziert:

| | |
|---|---|
| LAEST | (anstatt: LAESST, Rechtschreibfehler) |
| BTX-BRIEFEN | (BTX: nicht lexikalisiert) |
| DATENSCHUTZPROBLEME | (zerlegt: DATENSCHUTZ/PROBLEM) |
| BILDSCHIRMTEXT | (zerlegt: BILDSCHIRM/TEXT) |

II.2.2.2 Erweiterung des semantischen Regellexikons, des Fachlexikons und des CTX-Thesaurus

In den weiteren Schritten der Wörterbucharbeiten werden nur noch deskriptorrelevante Wortformen berücksichtigt. Als "Hilfsmittel" dazu lassen sich zu jedem Text bzw. zu jeder zu analysierenden Textmenge ein Grundformen-Index (sog. "Lemmatisierter Index") sowie ein KWIC-Index auf Wortformenebene erstellen. Im Grundformenindex sind die syntaktischen Homographen aufgelöst und diskontinuierliche Verbzusätze an das zugehörige Verbsimplex angefügt.

Im einzelnen schließen sich daran folgende Arbeitsschritte an:

- Abgleich des lemmatisierten Index mit dem Inventar des semantischen Lexikons und des Fachlexikons
  Dadurch werden die bereits (im Fachlexikon) inventarisierten eindeutigen Wörter und diejenigen semantisch mehrdeutigen Wörter, die schon bedeutungsdifferenziert sind, ausgefiltert; die übrigen "neu" hinzugekommenen werden intellektuell auf semantische Mehrdeutigkeit überprüft und ggf. entsprechend kodiert (vgl. Kap. II.3.3, Kap. II.3.4 und Kap. II.3.5).

- Dekomposition und Derivation
  Die nicht schon im Fachlexikon enthaltenen Wörter werden unter einem eher allgemeinsprachlichen und einem fachsprachlichen Gesichtspunkt bearbeitet. Es wird festgestellt, welche Komposita noch nicht zerlegt sind und zu welchen Lemmata sog. Derivate existieren, die im Fachlexikon fehlen (vgl. Kap. II.3.2).

- Semantische Relationierung
  Bei einer Anwendung von CTX-II (wie im vorliegenden Fall im Bereich Datenschutz) werden die "neuen" fachsprachlichen ein- und mehrwortigen Begriffe - soweit erforderlich - im CTX-Thesaurus mit den übrigen Fachbegriffen semantisch relationiert und entsprechend verzeichnet (vgl. Kap. II.3.6).

- Mehrwortige Begriffe
  Die Auswertung des KWIC-Index macht u.a. besonders häufig auftretende Kombinationen von Textwortformen sichtbar. Diese können daraufhin überprüft werden, ob sie als Mehrwortbegriff (Feste Wendungen im fachlichen Sinn) einzuordnen sind; gegebenenfalls werden hierzu im semantischen Regellexikon sowie im Fachlexikon entsprechende Markierungen vorgenommen. Der KWIC-Index stellt zugleich ein brauchbares Hilfsmittel zur intellektuellen Erarbeitung bzw. Ergänzung der Regeln des semantischen Lexikons dar.

In regelmäßigen Abständen muss (gleichsam in Umkehrung) das Fachlexikon sowie der CTX-Thesaurus mit den Analysewörterbüchern abgeglichen werden, da durch die Bildung von Derivationen und insbesondere durch semantische Relationierungen Begriffe eingehen können, die nicht notwendig bereits in den Analysewörterbüchern enthalten sind. Dies geschieht insbesondere unter dem Aspekt, dass bei einer natürlichsprachigen Anfrage beim Retrieval (vgl. Kap. 1.2.8) - zumindest auf die vorhandene Textbasis bezogen - von entsprechend konsistenten Wörterbüchern ausgegangen werden kann. Zur Wahrung der Konsistenz bei punktuellen Korrekturen werden Änderungsprotokolle angelegt, die in regelmäßigen Abständen aufgearbeitet werden.

Bezogen auf den Beispieltext waren folgende Wörter einer entsprechenden Kontrolle zu unterziehen:

| | |
|---|---|
| NEUE MEDIEN | (Feste Wendung: Fachlexikon, semantisches Lexikon) |
| MEDIUM | (mehrdeutiges Wort: Semantisches Lexikon, Differenzierungswörterbuch) |
| BTX | (Abkürzung: CTX-Thesaurus) |
| MANIPULATION | (morphosyntaktische Derivation) |
| BILDSCHIRMTEXT | (Kompositum: Fachlexikon) |

## II.3 Struktur und Inventar der Lexika

Da die verschiedenen Lexikonarbeiten sowohl aus der Sicht der möglichen Anwendungen von CTX wie auch in der Modellanwendung im Bereich Datenschutz eine wesentliche Rolle spielen, werden die Strukturen und das jeweilige Informationsinventar im folgenden ausführlicher beschrieben. Bezüglich einer ausführlichen Dokumentation der Algorithmen zur automatischen Sprachanalyse sei auf die Beschreibung SALEM 1980 verwiesen.

## II.3.1 Das morphosyntaktische Lexikon

### II.3.1.1 Struktur

Das morphosyntaktische Lexikon ist im wesentlichen ein Stammformenwörterbuch. Es wird zur Reduktion von Flexionsformen auf Grundformen und zur Ermittlung syntaktischer Angaben für die weitere Analyse eingesetzt. Es umfasst

- die Stämme (sog. "Identifikationswortlaute") der Hauptwortklassen, d.h. von Adjektiven, Verben und Substantiven;
- die Wortformen von sog. "Funktionswörtern" wie Adverbien, die nicht auf Adjektivstämme zurückzuführen sind, Präpositionen, Konjunktionen etc. Hier sind auch die Flexionsformen der Hilfsverben zugerechnet;
- eindeutige, nicht flektierbare, kontinuierliche Mehrwortbegriffe wie TREU UND GLAUBEN (nicht aber flektierbare Wortfolgen wie NATUERLICHE PERSON: diese werden als "Feste Wendungen" bezeichnet);
- Spezialeinträge zur Derivations-/Kompositumanalyse wie: Fugenmorpheme, Präfixe und Suffixe.

Zu jedem Paradigma der Hauptwortklassen werden alle zur Erzeugung bzw. Erkennung seiner zugehörigen Wortformen notwendigen "Stammformen" (Identifikationswortlaute) in das Lexikon aufgenommen. Ein Eintrag des morphosyntaktischen Lexikons besteht aus der Zeichenkette einer Stammform (ohne Endung) und syntaktischen und morphologischen Informationen wie Genus, Numerus sowie Angaben zu den möglichen Flexionsendungen; des weiteren sind syntaxorientierte Angaben zu Valenzen und Nebensatzanschlüssen kodiert. Je Grundform können mehrere Lexikoneinträge erforderlich sein, was bei Verben in der Regel auch der Fall ist. Für CTX ist vor allem die Kodierung der präpositionalen Valenzen von Bedeutung, da darauf die Identifikation der Komplexen P-Deskriptoren wie die Erkennung entsprechender Verbrelationen aufbaut.

Die entsprechenden Kodierkonventionen wurden am Sonderforschungsbereich Elektronische Sprachforschung (SFB 100) im Rahmen des MÜ-Systems entwickelt. Dort wurde auch der Großteil des Inventars des morphosyntaktischen Wörterbuches auf der Grundlage eines umfangreichen einsprachigen Wörterbuches (des "Deutschen Wörterbuchs" von G. Wahrig) mit rd. 80.000 Stichwörtern systematisch erfasst. Dieses Gesamtlexikon wurde v.a. auf Veranlassung und unter programmtechnischer Mitwirkung der Mitarbeiter des Forschungsprojekts JUDO für die Zwecke der automatischen Texterschließung verfügbar gemacht. Die in den Texten der Laborimplementierung auftretenden fehlenden Wörter wurden im Rahmen des Forschungsprojekts JUDO in eigener Verantwortung kodiert. Bei der Anwendung wurde kein spezifisches Analysewörterbuch erstellt, die Lexikondaten wurden vielmehr mit den vorhandenen kumuliert.

Zur effektiveren und effizienteren Erstellung der Substantiveinträge sowie der besonders umfangreichen Verbeinträge wurde im Rahmen des Forschungsprojekts JUDO ein Verfahren entwickelt, das eine sowohl zeitsparende als auch fehlerresistente Lexikonerstellung unterstützt: Unter Verwendung der Flexions- und Paradigmenkennung des Deutschen Wörterbuchs von G. Wahrig wurden Musterlisten zur Stammerzeugung und zur Flexion erstellt. Neue Wörterbucheinträge erhalten durch den Kodierer über eine Kennziffer einen Verweis auf das zugehörige Flexionsmuster; anschließend werden die erforderlichen Stammformen (Identifikationswortlaute)

und Flexionsangaben automatisch erzeugt. Syntaktische Angaben, wie z.B. Angaben zu Nebensatzanschlüssen und Präpositionalvalenzen, müssen allerdings weiterhin intellektuell angefügt werden.

Da nach den ursprünglichen Erfassungsrichtlinien des SFB 100 zu einem Paradigma der Wortklasse Verb im Höchstfall bis zu sechs Stammformen (z.B. bei einigen unregelmäßigen Verben) als Lexikoneinträge zu erstellen waren, stellt die Reduktion auf einen einzigen Eintrag eine nicht unerhebliche Ersparnis (allein an Schreibaufwand) dar. Es hat sich zudem in Tests gezeigt, dass gerade die Angaben zu den Flexionsendungen bei intellektueller Kodierung sehr fehleranfällig sind (solche Fehler führen dazu, dass der Reduktionsalgorithmus der Syntaxanalyse Textwortformen nicht erkennen kann). Gerade diese Angaben werden jetzt automatisch über Muster erzeugt, so dass eine bedeutende Fehlerquelle ausgeschaltet werden konnte.

II.3.1.2 Umfang

| PARADIGMEN | |
|---|---|
| SUBSTANTIVE | 68.000 |
| VERBEN | 12.000 |
| ADJEKTIVE | 12.500 |
| SONSTIGE | 2.300 |
| gesamt: | 94.000 |

| STAMMFORMEN (Identifikationswortlaute) | |
|---|---|
| SUBSTANTIVE | 77.000 |
| VERBEN | 42.500 |
| ADJEKTIVE | 12.700 |
| SONSTIGE | 2.500 |
| gesamt: | 135.000 |

(Stand: Februar 1982)

II.3.1.3 Kodierung

Aufgrund der systematischen Erweiterung des morphosyntaktischen Lexikons von ca. 18.000 (bei Beginn der 2. Projektphase Januar 1980) auf nunmehr ca. 135.000 Einträge verringerte sich der Kodieraufwand für "unbekannte Wörter" der Texte gegenüber den Tests des CTX-Systems in der ersten Projektphase beträchtlich. So ergaben sich für die Texte der Phase 2 bei ca. 133.000 laufenden Wortformen aus unterschiedlichen Textsorten folgende Zahlen:

| | |
|---|---|
| unbekannte Wörter insgesamt | 4407 |
| richtig identifizierte Wortzusammensetzungen | 3096 |
| Abkürzungen (bleiben als Substantiv erhalten) | 125 |
| Erfassungsfehler (werden korrigiert) | 124 |
| zu kodierende Flexionsformen | 1062 |

Da jedoch nicht die Flexionsformen, sondern Stammformen kodiert werden, war die Zahl der zu kodierenden Einträge geringer. Sie verteilte sich auf die einzelnen Wortklassen wie folgt:

| | |
|---|---|
| Substantive | 474 |
| Adjektive | 199 |
| Verben | 60 |
| Funktionswörter | 18 |
| insgesamt | 751 |

Da lediglich Einträge der Funktionswortklassen einen erhöhten Kodieraufwand erfordern (pro Eintrag ca. 1/2 Stunde), ein "normaler" Wörterbucheintrag für das morphosyntaktische Lexikon sich einschließlich Kontrolle und Rüstzeiten inzwischen jedoch in ca. 5 Minuten erstellen lässt, könnte die Kodierarbeit zur textorientierten Anpassung des morphosyntaktischen Lexikons - von der Überprüfung der unbekannten Wörter bis zur Erfassung m Dialog - von einem geübten Kodierer in zwei bis vier Wochen geleistet werden.

Durch eine (Re-)Integration der Ergebnisse der erkannten Wortzusammensetzungen könnten - wie erwähnt - eine weitere Reduktion des Kodieraufwands und zusätzliche Einsparungen im Rechenzeitverhalten erzielt werden. So wird z.B. in Zukunft die Überprüfung einmal richtig zerlegter Komposita entfallen.

### II.3.2 Morphosyntaktische Derivationsrelationen

Die (vorgegebene) Struktur des morphosyntaktischen Lexikons enthält keine Relationierungen zwischen den Stichwörtern, die "formal-inhaltlich" zusammengehören, aber unterschiedlichen Wortklassen angehören (den sog. Derivationen). Die Experimente des Modellsystems sollten jedoch zumindest die Möglichkeit einschließen, beim Retrieval systematisch wortklassenübergreifend vorzugehen, v.a. um den sog. "Recall", d.h. die Anzahl der relevanten Dokumente bei einer Suchanfrage, zu erhöhen.

Unter Derivationen werden hier Begriffe verstanden, die sich auf morphologischer Ebene voneinander ableiten lassen. Dies geschieht i.a. mittels Wortbildungsmorphemen, die an den Stamm eines Wortes angehängt werden und damit die Wortklasse bestimmen; Beispiele hierfür sind -bar,

-lich für Adjektive (z.B. ABLEIT-BAR) sowie -ung, -ion für Substantive (z.B. ABLEIT-UNG).

Im Rahmen der Thesaurusarbeiten im Projekt JUDO-DS wurde der allgemeine Begriff der Derivation unter bestimmten Aspekten eingegrenzt. Die morphologische Verwandtheit muss in jedem Falle berücksichtigt werden; es gibt jedoch Ableitungen, die sich im Laufe der Zeit in ihren Bedeutungen soweit voneinander entfernt haben, dass sie kaum noch als Begriffspaare empfunden werden (z.B. RECHT - GERECHT).

Unter diesen Gesichtspunkten wurde eine Systemkomponente entwickelt, die über eine intellektuell vorzunehmende Relationierung syntakto-semantischer Art Begriffspaare erstellt. Der Zweck dieser Komponente soll zunächst an einem Beispiel verdeutlicht werden: Wird z.B. in einer Suchanfrage der "Substantiv"-Deskriptor NACHWEIS verwendet, so ist es wahrscheinlich, dass auch Dokumente ausgegeben werden sollen, die das Verb NACHWEISEN (mit Textbelegen wie WIES ... NACH) und/oder die Adjektive NACHWEISBAR bzw. NACHWEISLICH enthalten.

II.3.2.1 Struktur

Entsprechend den deskriptorfähigen Wortklassen Verb, Substantiv und Adjektiv wurden drei Derivationsklassen aufgebaut. Alle drei Derivationsarten werden invertiert:

(1) DSV/DVS (Substantiv/Verb bzw. Verb/Substantiv)
SAMMLUNG - SAMMELN
AUSWAHL - AUSWAEHLEN

(2) DSA/DAS (Substantiv/Adjektiv bzw. Adjektiv/Substantiv)
MANIPULATION - MANIPULIERBAR
SELEKTIV - SELEKTION

(3) DVA/DAV (Verb/Adjektiv bzw. Adjektiv/Verb
SELEKTIV - SELEKTIEREN
MANIPULIEREN - MANIPULIERBAR

Diese derivierten Begriffe werden unter Angabe der jeweiligen Relation in den Thesaurus eingetragen und stehen damit (nach Umsetzung in den jeweiligen Thesaurus-Teil des IR-Systems) für eine erweiterte Recherche zur Verfügung. Dabei müssen in der Version CTX-II - d.h. bei Einschluss der semantischen Differenzierung von Homonymen - mehrdeutige Wörter in ihrer abgeleiteten Form auf die jeweils zutreffende Bedeutungsvariante abgebildet werden (vgl. Differenzierungswörterbuch Kap. II.3.4). Nicht immer handelt es sich dabei um 1:1-Entsprechungen. Folgende Fälle können auftreten (die Ziffern geben die jeweilige Bedeutungsvariante an):

(1) Einer Bedeutungsvariante kann entsprechend eine Derivation zugeordnet werden, z.B.

"Absonderung" besitzt zwei Bedeutungsvarianten:
ABSONDERUNG 01 (in der Bedeutung "Ausscheidung")
ABSONDERUNG 02 (in der Bedeutung "Isolation")

ebenso "absondern":
ABSONDERN 01 (in der Bedeutung "ausscheiden")

ABSONDERN 02 (in der Bedeutung "isolieren")

Man kann also relationieren:
ABSONDERUNG 01 DSV ABSONDERN 01
ABSONDERUNG 02 DSV ABSONDERN 02

(2) Nur ein Element des Derivationspaares ist mehrdeutig. In diesem Fall werden alle "passenden" Bedeutungsvarianten des mehrdeutigen Begriffs dem eindeutigen zugewiesen, z.B.

"Identität" besitzt zwei Bedeutungsvarianten:
IDENTITAET 01 (in der Bedeutung "Person")
IDENTITAET 02 (in der Bedeutung "Gleichheit")

"identisch" besitzt nur eine Bedeutung:
IDENTISCH (in der Bedeutung "gleich")

Man kann also relationieren:
IDENTITAET 02 DSA IDENTISCH

Für IDENTITAET01 ist eine solche Relation nicht angebracht.

(3) Beide Elemente eines Derivationspaares sind mehrdeutig, jedoch mit unterschiedlicher Anzahl der Bedeutungsvarianten. Hier werden alle "passenden" Differenzierungen einander zugeordnet, z.B.

"Reaktion" hat drei Bedeutungsvarianten:
REAKTION 01 (in der Bedeutung "chem. Wirkung")
REAKTION 02 (in der Bedeutung "Verhalten")
REAKTION 03 (in der Bedeutung "polit. Einstellung")

"reagieren" hat zwei Bedeutungsvarianten:
REAGIEREN 01 (in der Bedeutung "chem. Wirkung zeigen")
REAGIEREN 02 (in der Bedeutung "sich verhalten").

Man kann relationieren:
REAKTION 01 DSV REAGIEREN 01
REAKTION 02 DSV REAGIEREN 02

Hingegen hat REAKTION 03 keine verbale Entsprechung, jedoch lässt sich hier eine Derivationsrelation mit dem (eindeutigen) Adjektiv REAKTIONAER bilden.

### II.3.2.2 Umfang

Der Auf- und Ausbau der Derivationsrelation erfolgte stufenweise entsprechend dem Material der verarbeiteten Texte. Die folgende Statistik gibt eine Übersicht über die Anzahl der im Projektzeitraum erstellten Derivationspaare und ihre Verteilung auf die drei Derivationstypen (ohne

Invertierung).

| | | |
|---|---|---|
| Bestand insgesamt | 1518 Paare | 100,0% |
| DSV/DVS | 966 Paare | 63,8% |
| DSA/DAS | 334 Paare | 21,9% |
| DVA/DAV | 218 Paare | 14,3% |

(Stand: Februar 1982)

II.3.2.3 Kodierung

Ein interessanter Aspekt für eine mögliche weitgehende Automatisierung des Derivationsverfahrens ist, dass auf die Substantiv-Verb-Derivationen mehr als 2/3 aller Ableitungen entfallen. Dabei handelt es sich fast ausschließlich um regelmäßige Verb-Substantiv-Ableitungen (Verbstamm + EN / Substantiv mit Endung -UNG oder -ION).

Im Rahmen weiterer Arbeiten an den Derivationsrelationen wird die Behandlung der Partizipien (insbesondere der Partizipien II) noch zu diskutieren sein. Partizipien werden durch die Analyse auf das entsprechende Verb zurückgeführt, stehen also nicht als eigenständige Deskriptoren zur Verfügung; dabei ist es sehr häufig der Fall, dass Partizipien im Sprachgebrauch bereits Adjektivcharakter aufweisen und als lexikalisierte Form (wie z.B. BERECHTIGT) die Recherche erleichtern könnten. Diese Fälle wurden gegenwärtig dadurch gelöst, dass bei der Deskriptorerstellung attributiv gebrauchte Partizipien in ihrer Oberflächenrelisation rekonstruiert werden. Hierzu wird ein spezielles Hilfslexikon Verb/Partizip automatisch generiert. Dieses ist jedoch im folgenden nicht weiter ausgeführt.

II.3.3 Das semantische Regel-Lexikon

Die korrekte maschinelle Analyse eines Satzes schließt in der Anwendung CTX-II eine semantische Vereindeutigung der Textwörter ein. Unaufgelöste semantische Mehrdeutigkeiten erschweren naturgemäß die angemessene Weiterverarbeitung für eine wie auch immer geartete Verwendung, sei es die automatische Sprachübersetzung, die Thesauruserstellung oder die Indexierung für Information-Retrieval. Im Verfahren CTX ermöglicht erst die semantische Vereindeutigung die Nutzung eines Thesaurus und - damit verbunden - die Verringerung von Ballast bei der Recherche in der Informationsbank, d.h. die Erhöhung der Precision durch

- die eigentliche Bedeutungsdifferenzierung von Einzelbegriffen und
- die Zusammenführung von Einzelbegriffen zu so genannten Festen Wendungen (Mehrwortbegriffen).

Da eine intellektuelle (z.B. dialogorientierte) Vereindeutigung von Textwörtern bei der Dokumenterschließung - v.a. bei der Verarbeitung von Massendaten - kaum realistisch erscheint, müssen Verfahren entwickelt werden, die diesen Aufwand vermeiden oder zumindest - bei möglichst geringer Fehlerrate - entscheidend verringern. Es wäre vermessen, würde man hier auf Anhieb Fehlerlosigkeit und Vollständigkeit von einem automatischen Verfahren erwarten. Vielmehr war bereits für die Laboranwendung eine praxisnahe, d.h. auch mit vertretbarem Aufwand praktizierbare Losung dieses Problems intendiert, die auf mehreren, alternativen bzw. sich ergänzenden

Strategien aufbaut. Eine derartige Teillösung stellt die Verwendung der semantischen Analysekomponente dar, wie sie am Sonderforschungsbereich "Elektronische Sprachforschung" (SFB 100) in den letzten Jahren grundlegend entwickelt wurde. Dieses Analyseverfahren wurde anderweitig ausführlich dokumentiert (vgl. SALEM 1980). Hier soll es nur in Bezug auf die Verwendung im Verfahren CTX vorgestellt werden.

Dem Verfahren liegt ein spezifisches Merkmal- und Regellexikon zugrunde, das vorwiegend anwenderseitig zu pflegen ist. Mit Hilfe der darin vorhandenen Informationen wird die Zusammengehörigkeit von Worten und Wortgruppen eines zu analysierenden Satzes überprüft. Auf diese Weise können unzutreffende (d.h. innerhalb des Satzkontextes nicht passende) Bedeutungsvarianten ausgeschlossen und die richtigen Wortgruppen/Satzteile zusammengeführt werden. Dabei werden für die einzelnen Bedeutungsvarianten in der Anwendung CTX-II Kennziffern vergeben. Diese Kennziffern resultieren aus den in den Regeleinträgen vermerkten "Bedeutungsnummern", die wiederum aus den Angaben des Differenzierungslexikons (vgl. Kap. II.3.4) entnommen sind und auch bei der Differenzierung im CTX-Thesaurus usf. verwendet werden. Mithilfe dieser Kennzeichnung - die den Konventionen der Modellanwendung im Bereich Datenschutz entspricht, im Prinzip also extern auch durch andere Merkmale ersetzt werden könnte - wird bei mehrdeutigen Stichwörtern die Konsistenz aller betreffenden Lexika (unabhängig vom Fachgebiet) gesichert.

Das Ergebnis des automatischen semantischen Analyseverfahrens bildet den Abschluss der linguistischen Sprachanalyse. Auf dieses Ergebnis greift im Anschluss die Deskriptorensynthese zu. Sie erstellt aus den vorgefundenen Worten/Wortgruppen mitsamt ihren jeweiligen Vermerken Deskriptoren, die der Wiederauffindung des betreffenden Dokuments im Verlauf einer Recherche dienen (vgl. Kap. II.4).

#### II.3.3.1 Struktur

Das semantische Regellexikon enthält zwei Typen von Einträgen:

- Merkmaleinträge (nur für Substantive)
- Regeleinträge (für Substantive und Verben)

Aufgrund der in diesen Einträgen enthaltenen Informationen werden die verschiedenen Bedeutungsvarianten eines mehrdeutigen Wortlauts differenziert.

Merkmaleinträgekönnen zwei Arten von Angaben enthalten: zum einen semantische Merkmale (entsprechend der nachstehenden Liste; die Kombination mehrerer Angaben ist möglich).

| MERKMALE | BEDEUTUNGEN | BEISPIELE |
|---|---|---|
| ABS | abstrakt | Vision, Argument |
| KON | konkret | Haus, Auto |
| BEL | belebt | Baum |
| NBE | nicht belebt | Lied, Leiche |
| MEN | menschlich | Vater |
| TIE | tierisch | Fuchs |
| KOL | kollektiv | Bundestag, Verein |
| ORT | Ort | Brasilien, Wüste |

| | | |
|---|---|---|
| PUN | Punkt | Schnittstelle |
| LIN | linear | Gerade |
| FLA | Fläche | Seite |
| RAU | Raum | Saal, All |
| INS | Instrument | Hammer, Flöte |
| TER | Handlung | Spiel, Vortrag |
| ZEI | Zeitangabe | Stunde, März |
| WIS | Wissenschaftsgebiet | Physik |
| ZAL | zählbar | Person, Straße |
| NZA | nicht zählbar | Datenverarbeitung |

Zum anderen können Nomenklassifikationen vergeben werden, die sich an der Verträglichkeit mit bestimmten Präpositionen orientieren.

Regeleinträgeenthalten Angaben über die semantischen Beziehungen einer Bedeutungsvariante zu den sie umgebenden Wörtern, z.B. Valenzen (direkt, präpositional) und Informationen über Attribute bzw. Komplemente, teilweise (bei Festen Wendungen und unpersönlichem Verbgebrauch) einschließlich des Wortlauts.

Ein Eintrag besteht dabei aus folgenden Angaben:

WL1: Wortlaut des Stichworts

WL2: Wortlaut der Festen Wendung, die auf dem Stichwort basiert; liegt keine Feste Wendung vor, bleibt dieser Bereich normalerweise leer; er steht dann für spezielle ergänzende Angaben zur Verfügung.

IBED: Bedeutungsnummer des Eintrags für Abspeicherungszwecke, fortlaufend für Einträge mit identischem Stichwort und gleichem RECTYP.

RECTYP: Lexikonwortklasse des Stichworts
für Regeleinträge: RECTYP 66 = Wortklasse Substantiv
RECTYP 65 = Wortklasse Verb
für Merkmaleinträge: RECTYP 71 = Wortklasse Substantiv

REGEL: Informationen zum Stichwort

Im folgenden werden hierzu einige Kodierungsbeispiele (einschließlich einer kurz gefassten Regelerläuterung) gegeben:

| Eintrag | | Erläuterung |
|---|---|---|
| WL1 | HALTEN | |
| WL2 | | |
| IBED | 23 | |
| RECTYP | 65 | |
| REGEL | X=84,X=09,U=01 | Wenn dem Stichwort im aktuellen Satz ein Akkusativkomplement |

(84) und ein Präpositionalattribut nach "für" (09) zugeordnet werden können, dann liegt die Bedeutungsvariante "01" (des Stichworts im Differenzierungslexikon) vor.

```
WL1   HALTEN WL2
IBED 30
RECTYP 65
REGEL X=81,U=05.06      hat Nominativkomplement (81);
                        Bed.variante 05 und 06
```

---

```
WL1   MEDIUM WL2
IBED  21
RECTYP      66
REGEL X=09,X=84,U=01    hat Akkusativkomplement (X=84)
                        nach "von" (X=09); Bed.Variante 01

WL1   MEDIUM WL2
IBED  22
RECTYP      66
REGEL X=01,X=81,U=01.03         hat Nominativkomplement (X=81)
                                nach "als" (X=0l); Bed.Var. 01.03

WL1   MEDIUM WL
IBED  23
RECTYP      66
REGEL X=26,X=84,U=01.03         hat Akkusativattribut (X=84)
                                nach "zu" (X=26), Bed.Variante 01.03
WL1   MEDIUM WL2
IBED 01 RECTYP 71 REGEL INS  Das Stichwort besitzt das Merkmal "Instrument" (INS)
```

Als Disambiguierungsgrundlage für Einzelbegriffe, die in das semantische Regellexikon aufzunehmen sind, dient - wie bereits erwähnt - das Inventarverzeichnis für mehrdeutige Stichworte. Dabei ergeben sich derzeit allerdings einige Probleme:

Der syntakto-semantische Regelapparat kann der weit gefächerten, intellektuell differenzierten Variantenvielfalt nicht immer gerecht werden. So kommt es vor, dass Bedeutungsvarianten, die in dem Inventar als jeweils eigenständige Einträge mit eigener Bedeutungsnummer eingetragen sind, im Regellexikon in einem 'Sammeleintrag' bearbeitet werden müssen, da die Analysestrategie keine so weitgehende Disambiguierung erlaubt. Bei diesem Sammeleintrag werden dann die Bedeutungsnummern der verbleibenden Varianten als Bestandteil der Regel vermerkt.

Dabei sind aus technischen Gründen bisher maximal vier Varianten zulässig. Wird diese Grenze überschritten, gilt der betreffende Eintrag als nicht disambiguierbar. Er erhält in diesem Fall den Vermerk U=88. Schon eine Zusammenfassung von nur zwei Varianten bedeutet allerdings eine Einschränkung gegenüber der Bedeutungsdifferenzierung.

Eine zusätzliche Schwierigkeit stellen mehrdeutige Adjektive dar, da für sie noch keine Disambiguierungs-Regeln existieren. Sie machen jedoch zahlenmäßig nur einen geringen Anteil der Mehrdeutigkeiten aus (ca. 8%).

Ein weiteres Problem ergab sich aus der Verwendung der Merkmaleinträge. Da innerhalb der Regeleinträge auf semantische Merkmale des Attributs bzw. Komplements abgefragt werden kann, müssen alle evtl. vorkommenden Substantive mit semantischen Merkmalen markiert sein - also auch die eindeutigen. Es war daher notwendig, die im morphosyntaktischen Lexikon (vgl. Kap II.3.1) enthaltenen eindeutigen Substantive mit semantischen Informationen zu versehen. Hierzu wurden inzwischen - allerdings ohne Nachkontrolle - rd. 70.000 Einträge des in den 70-er Jahren kodierten 1. Saarbrücker Analysewörterbuches bezüglich der semantischen Merkmale in das Regellexikon übertragen.

Zur Vereinfachung der Kodierung häufig auftretender Strukturen bei Festen Wendungen werden eine Anzahl von so genannten 'Makros' verwendet, die zu diesem Zweck im Rahmen der Arbeiten des SFB "Elektronische Sprachforschung" entwickelt wurden. Diese Kodierungshilfen ermöglichen eine Reduzierung der Kodierzeit (pro Eintrag) um ca. 50%. Da sie mnemotechnisch benannt sind, ist durch sie außerdem eine leichtere (individuelle) Lesbarkeit und damit größere Korrekturfreundlichkeit der Einträge gewährleistet. Die Kodierung mit Hilfe von Makros erfolgt nach den folgenden Konventionen.

Der RECTYP der Festen Wendung richtet sich nach ihrem Stichwort. Es gilt die für die "normalen" Kodierungen übliche Regelung.

Die Bedeutungsnummern werden für das Stichwort (bei gleichem Rectyp) fortlaufend vergeben.

Der Wortlaut 1 (WL1) enthält das Stichwort, der Wortlaut 2 (WL2) den Wortlaut der Festen Wendung. Anstelle der Regel wird lediglich das jeweilige Makro-'Kürzel' mit Bezugswortlaut eingetragen. Die Regel enthält also den Wortlaut 1 des in Frage kommenden Makros und in direktem Anschluss daran, in spitze Klammern gesetzt, (hier dargestellt als: '( )') den lemmatisierten Wortlaut (bei Verben einschließlich Endung) des Bezugsworts innerhalb der Festen Wendung. Die einzelnen Angaben der Regeln haben folgende Bedeutung:

- X: Kasus des Attributs bzw. des Komplements
- D: Wortlaut des Attributs bzw. des Komplements
- G: Anzahl der Elemente der Präpositionalgruppe
- B1: Analysewortklasse des Attributs
- F1: Numerus des Komplements (02 = Plural, 01 = Singular)
- P1: Adjektivattribut

Die Teilregel "M=GAP*,U1" besagt, dass bei Zutreffen der Regel für den Wortlaut 1 ein zusätzlicher Eintrag (in der Analyse-Ergebnisdatei) mit dem Wortlaut 2 und gleicher Wortnummer erzeugt werden soll.

Alle Bestandteile (Einzelwörter) der Festen Wendung (z.B. VERLEIHEN, AUSDRUCK, AUSDRUCK VERLEIHEN) erhalten dabei die Markierung 'FS', um sie für die Deskriptorenerstellung zu sichern und von der weiteren Verarbeitung im semantischen Analyseverfahren auszuschließen.

Kodierungsbeispiele

Makro:
WL1: 1SUBADJ Substantiv mit Adjektiv
WL2: Attribut (erweiterungs
IBED: 77 fähig)
RECTYP: 66
REGEL: P1,D=:*,M=GAP*,Ul;

'Normalkodierung':
WL1: MEDIUM
WL2: NEUE MEDIEN
IBED: 11
RECTYP: 66
REGEL: Pl,D=NEU*,M=GAP*,Ul

'Makrokodierung':
WL 1: MEDIUM
WL2: NEUE MEDIEN
IBED: 11
RECTYP: 66
REGEL: ISUBADJ(NEU)
---
Makro:
WL1: 1SUBVON Substantiv mit Präpositional
WL: attribut nach "von" (ersetzt
RECTYP: 66 Genitiv, erweiterungsfähig)
REGEL: X=22,D=:*,M=GAP*,Ul;

'Normalkodierung':
WL1: EINHALTUNG
WL2: EINHALTUNG VON VORSCHRIFTEN
IBED: 14
RECTYP: 66
REGEL: X=22,D=VORSCHRIFT*,M=GAP*,Ul

'Makrokodierung':
Wil: EINHALTUNG
W12 EINHALTUNG VON VORSCHRIFTEN
IBED: 14
RECTYP: 66
REGEL: 1SUBVON(VORSCHRIFT)
---

Makro:
WL1: 1VRBAKK Verb mit Akkusativkomplement
WL2: (erweiterungsfähig)
IBED: 77

RECTYP: 65
REGEL: X=84,D=:*,M=GAP*,U1;

'Normalkodierung':
WL1: FASSEN
WL2: BESCHLUSS FASSEN
IBED: 11
RECTYP: 65
REGEL: X=84,D=BESCHLUSS*,M=GAP*,U1;

'Makrokodierung':
WL1: FASSEN
WL2: BESCHLUSS FASSEN
IBED: 11
RECTYP: 65
REGEL: 1VRBAKK(BESCHLUSS)

Für Feste Wendungen mit außergewöhnlicher bzw. bisher selten auftretender Struktur wurden aus Gründen der Effizienz noch keine Makros entwickelt. Sie müssen daher nach wie vor nach dem "herkömmlichen" Verfahren kodiert werden.

II.3.3.2 Umfang

| | | |
|---|---|---|
| Einträge insgesamt: | 76268 | 100,00% |
| | | |
| Wortklassen | | |
| Substantive: | 74487 | 97,66% |
| Verben: | 1737 | 2,28% |
| Makros: | 44 | 0,06% |
| | | |
| Eintragstypen | | |
| Merkmaleinträge: | 72496 | 95,05% |
| Regeleinträge: | 3728 | 4,89% |
| davon Einzelbegriffe: | 2914 | 3,82% |
| Feste Wendungen: | 814 | 1,07% |
| | | |
| Substantive | | |
| Einträge insgesamt: | 74487 | 100,00% |
| Merkmaleinträge: | 72496 | 97,33% |
| Regeleinträge: | 1991 | 2,67% |
| davon Einzelbegriffe: | 1469 | 1,97% |
| Feste Wendungen: | 522 | 0,70% |
| | | |
| Verben | | |
| Eintrage insgesamt: | 1737 | 100,00% |
| davon Einzelbegriffe: | 1445 | 83,19% |
| Feste Wendungen: | 292 | 16,81% |

(Stand: Februar 1982)

### II.3.3.3 Kodierung

Hier sind einige Erfahrungswerte festgehalten, die im Fall der Systemanwendung zu berücksichtigen sein werden:

Der reine Kodieraufwand für die Erstellung einer Regel zur Ermittlung einer Bedeutungsvariante ist 'unter normalen Bedingungen' (eingearbeitete Mitarbeiter) mit durchschnittlich 5 Minuten anzusetzen. Hinzu kommt bei der gegenwärtig erforderlichen On-line-Erfassung die Eintragszeit am Terminal, die in einem Time-Sharing-Betrieb weitgehend von der "Beschäftigung" des Rechners abhängt. Sie kann derzeit reduziert werden, wenn die Wörterbucheinträge in Zeiten geringerer Rechnerauslastung erfolgen. Im Falle einer CTX-Pilotanwendung wird die Möglichkeit eines anderen Erfassungsmodus zu prüfen sein. Verschiedene Konzepte wurden hierfür bereits entwikkelt.

Eine wesentliche Kodierungserleichterung wurde mit der Erstellung der Makros für Feste Wendungen (Mehrwortbegriffe) geschaffen. (Sie ist bei der Zeitabschätzung bereits berücksichtigt.) Die Liste der Makros lässt sich entsprechend den Bedürfnissen im Anwendungsfall beliebig erweitern. Die Verwendung von Makros für Einzelbegriffe erscheint indessen aufgrund der Vielzahl der Variationsmöglichkeiten nicht effizient.

Der Pflegeaufwand eines Lexikons ist im allgemeinen abhängig von Art und Umfang der notwendigen Korrekturen, sowie - in höherem Maße als bei Neueinträgen - von der Korrekturfreundlichkeit des Systems. Bei Änderungen einer Angabe im Eintrag bewegt sich der Korrekturaufwand um ca. 3 Minuten pro Eintrag (einschließlich On-line-Erfassung). Diese Zeit ließe sich bei gleichartigen (systematischen) Korrekturen durch Programmunterstützung reduzieren.

Unter Berücksichtigung anfallender Zusatzarbeiten (Korrekturen infolge der laufenden Inventaranpassungen, Dokumentation) kann davon ausgegangen werden, dass für die systematische Erstellung und Pflege von zehn- bis zwölftausend Einträgen des semantischen Lexikons etwa ein Mannjahr erforderlich ist. Mit dieser Anzahl der Einträge könnte etwa ein Fachgebiet (in der vorliegenden "Größe" des Datenschutzrechts) abgedeckt werden (einschließlich der dabei mit zu erfassenden allgemeinsprachlichen Bedeutungsvarianten (soweit nicht bereits vorhanden)).

Die Erstellung und Pflege des semantischen Lexikons könnte nach entsprechender Einarbeitungszeit (je nach Vorbildung etwa 2 - 4 Wochen) auch von Nichtlinguisten - z.B. auch von einem diplomierten Dokumentar - übernommen werden. Eine fachgebietsorientierte Vorbildung ist dabei sicherlich von Vorteil, desgleichen die Kenntnis linguistischer Grundbegriffe. Unerlässlich ist in jedem Fall eine fundierte Allgemeinbildung.

## II.3.4 Das Differenzierungswörterbuch

Im Rahmen der linguistischen Arbeiten stellte die Auflösung semantischer Mehrdeutigkeiten im Verfahren CTX-II einen Schwerpunkt dar.

Als mehrdeutig gilt im Sinne von CTX ein Wort, wenn unterschiedliche Begrifflichkeiten damit

bezeichnet werden. Dabei können mehrdeutige Begriffe im allgemeinen nur über den Kontext vereindeutigt werden. Die Bedeutungsvarianten können allgemeinsprachlicher Natur sein oder auch erst in einem bestimmten Fachgebiet als für eine Differenzierung relevant erscheinen. Da in Informationsdatenbanken, z.B. GOLEM oder TELDOK, bei einem Deskriptor nicht zwischen Groß- und Kleinschreibung unterschieden wird, muss bei gleicher äußerer Schreibweise (z.B. Laut - laut) auch eine wortklassenübergreifende Mehrdeutigkeit berücksichtigt werden.

Es ergeben sich daraus für eine semantische Disambiguierung folgende Arbeitskriterien:

(1) Das mehrdeutige Wort entstammt dem allgemeinsprachlichen Bereich und ist nur dort zu differenzieren.

(2) Das mehrdeutige Wort ist allgemeinsprachlich, hat aber eine fachgebietsspezifische Bedeutungsvariante.

(3) Das Wort ist fachgebietsspezifisch und soll in diesem Bereich bedeutungsdifferenziert werden.

(4) Die Mehrdeutigkeit ist wortklassenübergreifend.

Die mehrdeutigen Begriffe werden in einem ersten Schritt durch ihre Auflistung anhand der jeweils neu hinzukommenden Textmaterialien inventarisiert. Dabei werden zur Erkennung der Mehrdeutigkeiten sowohl allgemeinsprachliche als auch fachspezifische Wörterbücher herangezogen. Diese als mehrdeutig erkannten Wörter werden dann bedeutungsdifferenziert und in das Differenzierungswörterbuch eingetragen.

### II.3.4.1 Struktur

Im folgenden wird der Aufbau dieses Lexikons und seiner einzelnen Einträge dargestellt.

Ein Eintrag ist jeweils in zwei Komponenten untergliedert.

(1) Die erste Zeile enthält den Lemmanamen (die Grundform) sowie ein L zur Kennzeichnung. Ist der Eintrag nur einer Wortklasse zugehörig, wird die entsprechende Abkürzung angegeben (V = Verb, S = Substantiv, A = Adjektiv).

Beispiel:

| | | |
|---|---|---|
| L | S | DATUM |
| L | V | RICHTEN |
| L | A | GESETZLICH |

Ist der Eintrag wortklassenübergreifend, so bleibt die Spalte für die Kennzeichnung leer, die Wortklasse wird dann für jede Bedeutungsvariante gesondert angegeben.

(2) In den folgenden Zeilen des Eintrags sind die Bedeutungsvarianten angegeben. Jede Variante erhält (durchnummeriert) eine Zahl (01, 02, 03 ...) sowie eine Paraphrasierung und ein Beispielsyntagma zur Erläuterung. Zusätzlich werden Fachgebietsmarkierungen (z.B. zur Differenzierung von allgemeinsprachlichen und fachspezifischen Bedeutungsvarianten) sowie Gewichtungen vergeben. Die Gewichtungen dienen zur Auswahl der "wahr-

scheinlichsten" Variante, soweit alle anderen (linguistischen) Verfahren nicht zur automatischen Differenzierung geführt haben.

Beispiel:

| | | | |
|---|---|---|---|
| L | | | VERHALTEN |
| 01 | S | 9 | Benehmen: fragwürdiges Verhalten |
| 02 | V | 9 | refl.: s. benehmen: s. anständig verhalten |
| 03 | V | 0 | zurückhalten: den Schritt verhalten |

| | | | |
|---|---|---|---|
| L | S | | MEDIUM |
| 01 | | 9 | Mittel, Mittler: das Medium Fernsehen |
| 02 | | 2 | physikal. Substanz: löslich im Medium Wasser |
| 03 | | 0 | Verbindung zur Geisterwelt: ein gutes Medium |

Die Gewichtungen variieren von 0 bis 9, wobei 0 die niedrigste, 9 die höchste ist. Da die Analyse in jedem Falle syntaktisch differenziert, ist im obigen Beispiel bei einem identifizierten Substantiv automatisch auch die Bedeutungsvariante ermittelt. Sollte bei einem identifizierten Verb (oder Adjektiv-Partizip) die semantische Analyse keine geeignete Kontext-Regel anwenden können, so käme das statistische Verfahren zum Zuge, das im vorliegenden Falle die Bedeutung 02 bei Texten aus dem Datenschutz (da diese höher gewichtet ist als die Bedeutung 03) vorziehen würde. Der Cut-Off-Wert ist vom Anwender beeinflussbar. Die Gewichtung selbst stellt einen "Erfahrungswert" dar, der im Projektrahmen "simuliert" wurde, im konkreten Anwendungsfall jedoch auch statistisch-empirisch fundiert werden könnte. Für den Fall, dass mehrere Varianten über dem Cut-Off-Wert liegen, werden diese alle als (pseudo-) vereindeutigte Deskriptoren vergeben. Dies erlaubt die Beibehaltung der CTX-II-Konzeption.

Einen letzten Zusatz stellen die Fachgebietsmarkierungen dar, die allgemeinsprachliche Einträge (z.B. W = Welt) von fachgebietsspezifischen (z.B. R = Recht) unterscheiden.

Bei einer späteren Anwendung des Systems in einem anderen thematischen Bereich sind die Gewichtungsangaben und die fachgebietsspezifischen Kennzeichnungen entsprechend zu ändern. Ein mehrdeutiger Begriff wie "Zylinder" ist z.B. im technischen Fachbereich anders zu behandeln als im Textilwesen. Die Auflösung des Differenzierungswörterbuchs in mehrere, dem jeweiligen Fachgebiet entsprechende Mehrdeutigkeitslexika ist im Wesentlichen ein Problem der fachlichen Kompetenz, da nur ein Experte alle Nuancen eines Wortes auf seinem Gebiet beurteilen kann. Aus diesem Grunde wurden die Fachgebietsmarkierungen, die Gewichtungen und die Kontrolle der im Rechtswesen mehrdeutigen Wörter im Projekt JUDO-DS von juristisch ausgebildeten Mitarbeitern vorgenommen.

### II.3.4.2 Umfang

| | | |
|---|---|---|
| Anzahl der Lemmata | 935 | 100,00% |
| Anzahl der Varianten | 2945 | |

---

Wortklassenverteilung:

| | | |
|---|---|---|
| SUB | 493 | 52,83% |
| VRB | 310 | 33,05% |
| ADJ | 66 | 7,06% |
| wortklassenübergreifend | 66 | 7,06% |

---

Verteilung der Varianten:

| | | |
|---|---|---|
| Einträge mit 2 Varianten | 426 | 45,56% |
| Einträge mit 3 Varianten | 241 | 25,88% |
| Einträge mit 4 Varianten | 129 | 13,69% |
| Einträge mit 5 Varianten | 73 | 7,81% |
| Einträge mit 6 Varianten | 35 | 3,74% |
| Einträge mit 7 Varianten | 13 | 1,39% |
| Einträge mit 8 Varianten | 5 | 0,53% |
| Einträge mit 9 Varianten | 4 | 0,43% |
| Einträge mit 10 Varianten | 4 | 0,43% |
| Einträge mit 11 und mehr | 5 | 0,53% |

(Stand: Februar 1982)

### II.3.4.3 Kodierung

Die Arbeitsschritte bei der Erstellung eines Eintrages für dieses Wörterbuch sind:

- Erkennung der Mehrdeutigkeit
- semantische Differenzierung
- Vergabe der Gewichtung und der Fachgebietsmarkierung
- Erfassung
- evtl.. Korrektur

Der zeitliche Aufwand liegt pro Eintrag bei ca. 15 Minuten für alle Arbeitsschritte. Dabei ist die Fachgebietsmarkierung eine zusätzliche Informationshilfe für den Benutzer; ihre Angabe ist fakultativ und kann von den Bedürfnissen des Benutzers abhängig gemacht werden. Die Gewichtung wird bei der automatischen Sprachanalyse von SUSY nicht ausgewertet; sie ermöglicht jedoch heuristische Strategien zur Lösung der nach der semantischen Analyse verbliebenen Restmehrdeutigkeiten. Beim Wechsel eines Fachgebiets behalten die Einträge des Differenzierungswörterbuchs ihre Gültigkeit bis auf die Gewichtungsangaben.

## II.3.5 Das Lexikon fachspezifischer Deskriptoren

Das Lexikon fachspezifischer Deskriptoren wird textorientiert aufgebaut. Es enthält kumulativ die in der verarbeiteten Textmenge auftretenden deskriptorfähigen Begriffe mit zusätzlichen fachspezifischen Informationen. Das in diesem Wörterbuch gesammelte Wortmaterial bildet die Basis für den CTX-Thesaurus. Dabei kann die vorgenommene fachspezifische Klassifikation von Begriffen in doppelter Weise zur Unterstützung bei der Deskriptorerstellung herangezogen werden: Einerseits könnte die Deskriptorenliste beliebig durch eine Beschränkung auf fachspezifisch

relevante Begriffe gekürzt werden; im einfachsten Fall kann dadurch die Vergabe irrelevanter Deskriptoren, die über die Sprachanalyse nicht ausgefiltert werden konnten (z.B. Modalverben, aber auch fachspezifische Allerweltsbegriffe wie "Datenschutz" im Informationsrecht), ausgeschlossen werden. Weiterhin kann die Disambiguierung semantischer Mehrdeutigkeiten durch Auswertung der Fachgebietsmarkierung unterstützt werden. Voraussetzung dafür ist allerdings eine wesentlich umfangreichere Differenzierung der Fachgebietskennzeichnung, als sie im vorliegenden Projektzeitraum für die Laboranwendung geleistet werden konnte.

#### II.3.5.1 Struktur

Das Inventar dieses Lexikons bilden v.a. die fachgebietsspezifischen Deskriptoren. Darunter werden die deskriptorfähigen Einzelbegriffe bzw. Komposita und die sog. Festen Wendungen mit ihren Teilbegriffen verstanden.

Als deskriptorfähige Begriffe gelten im System CTX zunächst alle in den zu beschreibenden Texten auftretenden Textwörter der Wortklassen Substantiv, Verb und Adjektiv, sowie die von Adjektiven abgeleiteten Adverbien (die sog. "Einfachen Deskriptoren"). Semantisch mehrdeutige Wörter werden sowohl in ihrer unvereindeutigten Form (als sog. "Formaldeskriptoren") als auch mit allen ihren möglichen vereindeutigten Varianten aufgenommen. (Dies ermöglicht eine Freitextsuche auch ohne Berücksichtigung der Disambiguierung. Allerdings ist keine semantische Relationierung zwischen Formaldeskriptoren möglich).

Eine Kompositummarkierung erhalten in diesem Lexikon - abweichend vom üblichen Verständnis des Begriffs Kompositum - diejenigen Begriffe, die semantisch gesehen mindestens ein Zerlegungselement (Teilwort) enthalten, das im Sinne der Dokumentdeskribierung deskriptorrelevant ist (z.B. wird STAATSANWALTSCHAFT als Kompositum markiert mit dem Teilwort STAATSANWALT). Einem Kompositum werden neben seinen deskriptorfähigen Bestandteilen alle sinnvollen Kombinationen dieser Bestandteile (ggf. in vereindeutigter Form) zugeordnet. Auch diese werden als deskriptorfähige Begriffe inventarisiert.

So könnte z.B. das Kompositum BUNDESDATENSCHUTZGESETZ zerlegt werden in

| | |
|---|---|
| BUND1 | BUNDESGESETZ |
| DATUM1 | DATENSCHUTZ |
| GESETZ1 | DATENGESETZ |
| SCHUTZ | SCHUTZGESETZ |

Die Kombination BUNDESSCHUTZ hingegen erscheint als nicht sinnvoll. Sie ist zudem in der Fachsprache "Recht" nicht belegt und wird verworfen.

Die aus den Einfachen Deskriptoren generierten Komplexen Deskriptoren werden nur in Ausnahmefällen in das fachspezifische Lexikon aufgenommen, auch um den Umfang des Lexikons durch "Augenblickskonstruktionen" nicht zu sehr auszuweiten. Eine Ausnahme bilden die lexikalisierten "Festen Wendungen" bzw. Komposita und die mit diesen relationierten Komplexen Deskriptoren.

Als "Feste Wendungen" werden "registerfähige" (d.h. in Registern zur Fachliteratur auftretende) Mehrwortbegriffe verstanden, die in ihrer geläufigen Form als Suchbegriffe zur Verfügung ge-

stellt werden sollen. Darunter fallen zunächst Begriffe, deren Bedeutung sich nicht unmittelbar aus der Bedeutung ihrer Bestandteile erschließen lässt (z.B. NATUERLICHE PERSON). In einem weiteren Sinn werden darunter auch mehrwortige Begriffe gefasst, die in einem Fachgebiet häufig gebraucht wurden und so in ihrer Gesamtheit als Fachbegriff gelten. Bis auf ganz wenige Ausnahmen sind Feste Wendungen semantisch eindeutig (Ausnahme etwa: OEFFENTLICHES AMT, bei dem sich die unterschiedlichen Bedeutungen aus dem unterschiedlichen Gebrauch von AMT ableiten und durch OEFFENTLICH gerade nicht vereindeutigt werden).

Jeder dieser deskriptorfähigen Begriffe erhält als zusätzliche Information eine sog. "Fachgebietsmarkierung" und eine "Relevanzmarkierung".

(a) Fachgebietsmarkierung

Es wird (bezogen auf die Laboranwendung "Datenschutzrecht") markiert, ob der Begriff dem Spezialgebiet Datenschutzrecht (D) und/oder dem Fachgebiet EDV (E) und/oder dem Gebiet Recht (R) und/oder der Allgemeinsprache (W=Welt) zuzuordnen ist. Die letztere Markierung erhalten all jene Begriffe, die sich nicht einem der drei erstgenannten Spezialgebiete zuordnen lassen. Die Fachgebietsmarkierung soll folgende Funktionen unterstützen:

- die Vereindeutigung mehrdeutiger Wörter bei der Indexierung oder beim ("NATURA-")Retrieval;
- die Steuerung der Deskriptorerstellung durch Einschränkung auf vorgegebene Markierungen;
- die Unterstützung fachspezifischer Textsortenuntersuchungen.

Die vorgenommene fachspezifische Differenzierung im Datenschutzrecht erwies sich als relativ grob; es lassen sich jedoch Anwendungsgebiete denken, die aufgrund herkömmlicher Begriffsklassifikation einen viel versprechenden Einsatz dieser Information hei der Verarbeitung von Texten aus einem zum Teil recht heterogenen Datenbestand (Pool) nahe legen (z.B. könnte bei der Patentdokumentation die Internationale Patentklassifikation integriert werden).

(b) Relevanzmarkierung

Die inventarisierten Begriffe erhalten eine Zusatzinformation über ihre Relevanz für die inhaltliche Erschließung eines Dokuments. Diese Markierung kann analog der Gewichtungsfunktion bei einer "Filterung" zur Vergabe/Nichtvergabe eines Begriffes als Deskriptor ausgewertet werden. Sie unterscheidet, ob ein Begriff

- innerhalb des Fachgebiets von hoher Bedeutung ist,
- keinen selbständigen juristischen Aussagegehalt hat,
- als Stoppwort klassifiziert werden soll (z.B. Modalverben SOLLEN, MUESSEN, DUERFEN),
- nichtdeskriptorfähiger Bestandteil eines Kompositums ist (z.B. SCHAFT in STAATSANWALTSCHAFT); dies dient der systematischen Kontrolle des Thesaurus).

Beispiele für Einträge im fachspezifischen Lexikon:

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| SUB | | DE | 1 | | 5 | 01 DATUM | 1 |
| SUB | F2 | DE | 5 | P | 17 | ABGELEITETE DATEN | 01 |
| SUB | K2 | D | 5 | | 13 | DATENSAMMLUNG | 1 |
| SUB | F2 | D | 4 | | 11 | NEUE MEDIEN | 11 |

Dabei bedeuten die Informationen in den Bereichen 1-8

1: Wortklasse (SUB = SUBSTANTIV)
2: Kennzeichen: F2 Feste Wendung mit 2 Bestandteilen
K2 Kompositum mit 2 Bestandteilen
3: Fachgebietsmarkierung: D Datenschutzrecht
E EDV
4: Relevanzmarkierung: zwischen 0-6 mit zunehmenden Gewicht
5: Zusatzinformation: P Plurale tantum
6: Anzahl der Zeichen des Wortlauts (Leerstellen zwischen den Wörtern werden mitgezählt)
7: Wortlaut (mit Bedeutungsnummer z.B. 01)
8: Großschreibungsmarkierung: 1 = Wort beginnt mit Großbuchstaben

II.3.5.2 Umfang

| | | |
|---|---|---|
| Gesamtanzahl der Einträge: | 12.416 | 100,00% |
| Art der Einträge: | | |
| Deskriptorfähige Einzelbegriffe | 7.590 | 62,13% |
| Komposita | 3.428 | 27,61% |
| Feste Wendungen | 850 | 6,85% |
| sonstige mehrwortige Begriffe | 548 | 4,41% |
| Verteilung auf Wortklassen: | 12.416 | 100,00% |
| SUB (Substantiv) | 8.420 | 67,82% |
| VRB (Verb) | 2.139 | 17,23% |
| ADJ (Adjektiv) | 1.313 | 10,58% |
| PSE (Pseudo-Wortklasse) | 34 | 0,27% |
| ADV (Adverb) | 120 | 0,97% |
| EIG (Eigenname) | 259 | 2,09% |
| ABK (Abkürzung) | 131 | 1,04% |
| Verteilung nach Fachgebietsmarkierung: | 12.416 | 100,00% |
| D (Datenschutzrecht) | 1.021 | 8,22% |
| E (EDV-Bereich) | 186 | 1,50% |

| | | | |
|---|---|---|---|
| R | (allg. Recht) | 1.538 | 12,39% |
| W | ("Rest"-Welt) | 4.479 | 36,08% |
| DE | (Kombination D/E) | 1.338 | 10,78% |
| DR | (Kombination D/R) | 663 | 5,34% |
| DW | (Kombination D/W) | 800 | 6,44% |
| ER | (Kombination E/R) | 4 | 0,03% |
| EW | (Kombination E/W) | 109 | 0,88% |
| RW | (Kombination R/W) | 1.514 | 12,19% |
| DER | (Kombination D/E) | 20 | 0,16% |
| DEW | (Kombination D/E/W) | 327 | 2,63% |
| DRW | (Kombination D/R/W) | 415 | 3,36% |
| ERW | (Kombination E/R/W) | 0 | 0,00% |

---

(Stand: Februar 1982)

II.3.5.3 Kodierung

Das Lexikon fachspezifischer Deskriptoren wurde sukzessive an Hand des bearbeiteten Textmaterials von fachspezifisch (hier: juristisch) ausgebildeten Mitarbeitern aufgebaut.

Zur Auswertung der deskriptorfähigen Einzelbegriffe wurden aufgrund der Analyseergebnisse lemmatisierte Wortlisten erstellt, sortiert nach Wortklassen; die sog. Stoppwörter sind dabei bereits automatisch ausgefiltert. Nach einem maschinellen Abgleich mit dem schon vorhandenen Vokabular mussten dann nur noch die "neuen" Wörter in Bezug auf ihre Relevanz und ihre Fachzugehörigkeit untersucht werden.

Für die Zerlegung der Komposita in ihre sinnvollen Bestandteile kann der Zerlegungsalgorithmus des Analysesystems ausgenutzt werden: Der Kodierer erhält eine Liste mit Komposita und Zerlegungsvorschlägen, die auf falsche bzw. unzureichende Zerlegungen hin überprüft werden müssen. Auch die Festen Wendungen wurden nicht systematisch, sondern unter Heranziehung von Fachlexika oder entsprechenden Registern, mit Bezug auf das Inventar der bearbeiteten Texte, erstellt. Dabei erwies sich als zeitintensivster Faktor die Erkennung von mehrwortigen Sequenzen als Fachbegriff. Um (auch dem Fachmann) diese Arbeit zu erleichtern, können sog. Keyword-in--Context-Listen erstellt werden; Es handelt sich dabei um alphabetisch sortierte Listen von Textwörtern mit ihrer textuellen Umgebung; dabei ist die Umgebung frei zu definieren (entweder summarisch, d.h. durch eine festgesetzte Anzahl von Wörtern vor bzw. nach dem Stichwort; oder linguistisch orientiert, d.h. das Stichwort mit dem Satz, in dem es vorkommt). Dabei wird davon ausgegangen, dass durch solche Listen überdurchschnittlich häufig gebrauchte Sequenzen sichtbar werden und sich somit als fachliche Mehrwortbegriffe anbieten.

Da das Wortmaterial dieses Lexikons bedeutungsdifferenziert wurde, musste ein Abgleich mit dem Differenzierungswörterbuch durchgeführt werden, um der Konsistenz der Wörterbücher des Gesamtsystems zu genügen. Auch diese zeitaufwendige und fehleranfällige Kodierarbeit kann inzwischen durch entsprechende Hilfsroutinen unterstützt werden.

In diesem Zusammenhang konnten aufgrund des differenzierten Sprachanalyseverfahrens umfangreiche Hilfsmittel zur Erstellung dieses Lexikons zur Verfügung gestellt werden.

## II.3.6 Der CTX-Thesaurus

Der in dem fachspezifischen Lexikon aufbereitete (Fach-)Wortschatz bildet die Grundlage für den CTX-Thesaurus.

Die Begriffsbeziehungen im Thesaurus bauen auf semantisch eindeutigen Begriffen auf. Die Eindeutigkeit dieser Begriffe ist allgemeinsprachlich und damit fachgebietsunabhängig. Daraus ergeben sich für die Funktion und Struktur dieses Thesaurus Besonderheiten, die von den Begriffsbestimmungen in den "Richtlinien für die Erstellung und Weiterentwicklung von Thesauri (DIN 1463)" abweichen. Dort heißt es: "Nach seiner Funktion ist ein Thesaurus ein Mittel der terminologischen Kontrolle. (...) Nach seiner Struktur ist ein Thesaurus ein kontrolliertes, dynamisches Vokabular von bedeutungsmäßig und generisch verbundenen Termini, das umfassend einen spezifischen Fachbereich abdeckt." (DIN 1463, 1976; Unterstreichungen vom Verf.)

Durch die Vernetzung von fachgebietsunabhängig-eindeutigen Begriffen lassen sich die aufgestellten Beziehungen des CTX-Thesaurus unverändert auf andere Fachgebiete übertragen; somit ist es möglich, verschiedene Fachgebiete mit einem einzigen Thesaurus zu betreuen. Thesaurusarbeit für ein "neues" Fachgebiet fällt also nicht im Bereich der Änderung bzw. Anpassung von Beziehungen an, sondern in der Ergänzung um das spezifische Fachvokabular.

Für die unterschiedlichen Bezeichnungen eines Begriffs gibt es im CTX-Thesaurus keine ausgezeichnete Benennung. Damit ist dieser Thesaurus auch kein Instrument zur terminologischen Kontrolle im Sinne von Vorzugsbenennungen. Es handelt sich vielmehr um eine Sammlung von Wortfeldern, die durch Beziehungen strukturiert sind, welche in Anlehnung an DIN 1463 gewählt wurden. Allerdings ließe sich ein derartiges Instrumentarium zur terminologischen Kontrolle leicht implementieren, da eine Erweiterung der Relationen im System vorgesehen ist. Aufgrund der Orientierung an Online-Retrievalsystemen im Freitextbereich erschien diese Erweiterung, die bei einer intellektuellen Indexierung mit Bezug auf gedruckte Dienste durchaus üblich ist, in der Modellentwicklung nicht wichtig. Mit der semantischen Relationierung wird eine Brücke vom Indexierungsvokabular des Textes zum Retrievalvokabular des Benutzers geschlagen. Als Suchbegriffe werden die in den Texten verwendeten Wörter morphosyntaktisch normiert und semantisch vereindeutigt zugeteilt; über die semantischen Beziehungen stehen dem Benutzer auch semantisch verwandte Begriffe für die Recherche zur Verfügung. "Terminologische Kontrolle des Vokabulars" findet also in dem Sinn statt, dass inhaltliche Beziehungen zwischen den Wörtern intellektuell hergestellt werden.

Systembezogen erfüllt der CTX-Thesaurus eine zweifache Funktion: er unterstützt den Benutzer beim Retrieval in der oben angegebenen Weise (Recherchefunktion). Die aufgestellten Beziehungen zwischen Begriffen können aber auch unmittelbar zur Verbesserung der Texterschließung benutzt werden (Indexierungsfunktion).

### II.3.6.1 Struktur

Im CTX-Thesaurus sind die deskriptorfähigen Begriffe durch Relationen untereinander vernetzt. Die Begriffe werden in der Regel in ihrer vereindeutigten Grundform aufgeführt; nur in Ausnahmefällen, wenn die entsprechenden Wortformen geläufiger erscheinen (z.B. DATEN vs. DATUM im Sinne von Merkmalsangaben), werden auch Wortformen (hier: Nominativ-Plural-Form)

zur Relationierung zugelassen.

Die Relationen lassen sich in 5 Typen klassifizieren:

(a) Äquivalenzrelationen
(b) Hierarchierelationen
(c) Assoziationsrelationen
(d) Morphosyntaktische Relationen
(e) Fachspezifische Relationen

Ein Eintrag im CTX-Thesaurus hat die Form T1 R T2 , wobei der Begriff T1 (Ausgangsbegriff) zum Begriff T2 (Relatum) in der Relation R steht.

Eine Relation R heißt invertierbar, wenn es eine Relation R' gibt, mit T2 R' T1 für alle Paare (T1,T2), die in der Relation R stehen. Das Relationenkennzeichen ist durch ein dreibuchstabiges Kürzel (unterstrichen) dargestellt.

Beispiel: GESETZ UNT RECHTSVORSCHRIFT bedeutet:
der Begriff "Gesetz" ist Unterbegriff zum Begriff "Rechtsvorschrift".

zu (a): Äquivalenzrelationen

Unter die Äquivalenzrelationen sind die strenge Synonymie-Relation sowie eine Reihe weiterer eingeschränkter Synonymie-Relationen gefasst.

- Strenge Synonymierelation

Definition: Ausgangsbegriff und Relatum sind streng bedeutungsgleich und keiner der weiter unten beschriebenen eingeschränkten Synonymierelationen zuzuordnen.

Beispiele:
ABDRUCK1 SYN KOPIE
ABGEORDNETENHAUS SYN PARLAMENT
UNZUTREFFEND SYN UNRICHTIG

- Mehrwortrelationen

Definition: Komposita, Feste Wendungen und andere mehrwortige Begriffe sowie Komplexe Deskriptoren sind über Mehrwortrelationen miteinander verknüpft, wenn sie aus den gleichen Wortelementen aufgebaut sind. Der 1. Buchstabe des Kürzels (S=Synonymie). besagt, dass es sich bei dieser Relation um eine Synonymierelation handelt; der 2. und 3. Buchstabe bezeichnen die Art des Ausgangsbegriffs bzw. Relatums: (K = Kompositum, F = Feste Wendung, M = mehrwortiger Begriff, A = Komplexer Deskriptor (Adjektivrelation), C = Komplexer Deskriptor (Konjunktion), G=Komplexer Deskriptor (Genitivrelation), P=Komplexer Deskriptor (Präpositionalrelation).

Beispiele:
DOKUMENTATIONSZUSTAND SKM ZUSTAND DER DOKUMENTATION
DOKUMENTATIONSZUSTAND SKG ZUSTAND G DOKUMENTATION

| | |
|---|---|
| GEWERBEBETRIEB | BKA GEWERBLICHER BETRIEB |
| AMTSBLATT FÜR BERLIN | SFP AMTSBLATT P BERLIN |

- Relation orthografischer Varianten

Definition: Ausgangsbegriff und Relatum unterscheiden sich nur geringfügig in ihrer Schreibweise. Singular- und Pluralformen (die dann z.B. aufgenommen werden, wenn die Pluralform als Schlagwort üblich ist) desselben Begriffs werden unter dieser Relation aufgeführt.

Beispiele:

| | |
|---|---|
| BUNDESDATENSCHUTZGESETZ | SYS BUNDES-DATENSCHUTZGESETZ |
| PERSONENBEZOGENE DATEN | SYS PERSONENBEZOGENES DATUM |
| SCHADENERSATZ | SYS SCHADENSERSATZ |

Der Unterschied in der Schreibweise betrifft z.B. das Auftreten mit und ohne Bindestrich, mit und ohne Fugen-S bei Komposita sowie die Varianten Plural/Singular.

Die Relation orthografischer Varianten und die Mehrwortrelation werden v.a. zu systemseitigen Test- und Dokumentationszwecken sehr fein differenziert. Bei der Umsetzung in die IR-Systeme GOLEM und TELDOK wurden sie unter der "strengen Synonym-Relation" gefasst; obwohl es sich im engeren Sinn (besonders bei der Relationierung Komplexer Deskriptoren) nicht immer um Synonyme handelt.

- Quasisynonymierelation

Definition: Der Ausgangsbegriff hat "ähnliche" Bedeutung wie das Relatum. Das Maß der Ähnlichkeit kann anhand einer Skala mit 9 Stufen (in Zehnerschritten von 10 bis 90) angegeben werden. Für Recherchezwecke (GOLEM-Variante) wurden diese zu zwei Stufen (0 und 5) zusammengefasst.

Beispiele:

| | |
|---|---|
| BEFUGTER | QUA 90 BERECHTIGTER |
| DATENVERARBEITUNG | QUA 50 DATENVERAENDERUNG |
| AUFZEICHNUN1 | QUA 20 UNTERLAGE |
| ZUR KENNTNIS BRINGEN | QUA 10 ZULEITEN |

- Übersetzungsrelation

Definition: Die Übersetzungsrelation stellt die Beziehungen zwischen deutschsprachigen Begriffen und fremdsprachigen Äquivalenten her.
Ausgangsbegriff ist der deutsche Begriff, Relatum ist das fremdsprachige Äquivalent. Die 2. Stelle des Kürzels bezeichnet die Sprache des Ausgangsbegriffs, die 3. Stelle bezeichnet diejenige des Relatums (D = Deutsch, E = Englisch, F = Französisch).

Beispiele:

| | |
|---|---|
| DATENSCHUTZ | BDE DATA PROTECTION |
| DATENSCHUTZ | BOF PROTECTION DE DONNEES |
| DATA SECURITY | BEF SECURITE DES DONNES. |

zu (b): Hierarchierelation

- Abstraktionsrelationen (Ober-/Unterbegriff)

Definition: In der Oberbegriffsrelation (OBR) ist der Ausgangsbegriff Oberbegriff zum Relatum. Die Relation ist invertierbar (inverse Relation: Unterbegriffsrelation (UNT)). Der Unterbegriff stellt eine Spezialisierung des Überbegriffs dar.

| Beispiele: | | | |
|---|---|---|---|
| | BUNDESBEHÖRDE | OBR | BUNDESARBEITSAMT |
| | DATENDOKUMENTATION | UNT | DOKUMENTATION |

Anhaltspunkte für das Vorliegen einer Oberbegriffsbeziehung sind:

- Der Unterbegriff kann aus dem Oberbegriff durch Hinzufügen eines Attributs, durch die Bildung einer Festen Wendung oder durch Wortzusammensetzung gebildet sein.

  Beispiele hierfür:

| | |
|---|---|
| DATEI | OBR VERSCHLOSSENE DATEI |
| BUNDESAMT | OBR BUNDESAMT FUER VERFASSUNGSSCHUTZ |

- Aus textuellen Formulierungen wie z.B. § 7 Abs. 1 Satz 1 BDSG: "Die Vorschriften dieses Abschnitts gelten für Behörden und sonstige öffentliche Stellen des Bundes, der bundesmittelbaren Körperschaften, ..." kann gefolgert werden, dass OEFFENTLICHE STELLE Oberbegriff von BEHOERDE ist. Ein anderes Beispiel: § 39 Abs. 1 BDSG enthält die Formulierung "...Gesellschaften und andere Personenvereinigungen ...", woraus man schließen kann, dass PERSONENVEREINIGUNG Oberbegriff zu GESELLSCHAFT ist.

- Partitive Relationen (Teil/Ganzes)

Definition: Bei der Relation GAN (Ganzes) enthält das durch den Ausgangsbegriff Bezeichnete das durch das Relatum Bezeichnete.
Die Relation TEI (Teil) ist invers zu GAN.

| Beispiele: | | | |
|---|---|---|---|
| | DATEI | TEI | INFORMATIONSSYSTEM |
| | DATENBANK | GAN | DATEI |
| | DATENSCHUTZMASSNAHME | TEI | DATENSCHUTZ |

Nun ist z.B. nicht jede Datei Teil eines Informationssystems oder z.B. nicht jede Stadt Teil eines Bundeslandes. Deshalb wird die Definition der partitiven Relation GAN folgendermaßen präzisiert: Objekte des durch den Ausgangsbegriff bezeichneten Objekttyps enthalten i.a. Objekte von der Art des durch das Relatum bezeichneten Objekttyps.

zu (c): Assoziationsrelationen

- Assoziationsrelation

Definition: Der Ausgangsbegriff ist frei assoziiert zum Relator und ist mit diesem nicht durch eine der anderen beschriebenen Relationen verknüpft. Für Retrievalzwecke werden die 10 unterschiedlichen Gewichtungen zu 2 Klassen (0 und 5) zusammengefasst.

| Beispiele: | | |
|---|---|---|
| | EINWOHNERMELDEAMT | ASS 80 MELDEPFLICHT |
| | AERZTLICH | ASS 80 MEDIZINISCH |
| | AERZTLICHES GEHEIMNIS | ASS 50 VERSCHWIEGENHEIT |
| | BEHAELTNIS | ASS 20 RECHENZENTRUM |
| | AUSKUNFT1 | ASS 20 KENNTNIS |
| | ANZEIGE1 | ASS 20 DATENBANKREGISTER |

Kriterien und Anhaltspunkte für eine Assoziationsbeziehung zwischen zwei Begriffen sind etwa, dass sie in enger Beziehung zueinander stehen, zum selben engen Thema gehören und, vor allem, dass zu vermuten ist, dass Texte (Dokumente), die für den einen Begriff relevant sind, auch für den assoziierten Begriff relevant sind. (V.a. in diesem Bereich lassen sich statistische Verfahren vorstellen, die eine intellektuelle Assoziierung mit Vorschlägen unterstützen könnten.)

Assoziationsbeziehungen enthalten u.a. die Relationstypen Ursache - Wirkung, Tätigkeit - Ergebnis der Tätigkeit, Tätigkeit - Objekt der Tätigkeit, Tätigkeit - Mittel der Tätigkeit (instrumentelle Relation), die Relation 'ist Information über' u.a. Von einer derartigen expliziten Aufgliederung der Assoziationsrelation wurde aus pragmatischen Gründen abgesehen.

- Antonymrelation (Gegenbegriff)

Definition: Der Ausgangsbegriff ist Gegenbegriff zum Relatum.

| Beispiele: | | |
|---|---|---|
| | BEFUGTER | GEG UNBEFUGTER |
| | AKTIVITAET | GEG PASSIVITAET |

Die Gegenbegriffsrelation steht in gewisser Konkurrenz zur Nebenbegriffsrelation. In Zweifelsfällen wird das Begriffspaar der Gegenbegriffsrelation zugeordnet.

- Kohyponymrelation (Nebenbegriff)

Definition: Der Ausgangsbegriff ist Nebenbegriff zum Relatum.

| Beispiele: | | |
|---|---|---|
| | DATENWEITERGABE | NEB DATENEINGABE |
| | DATENWEITERGABE | NEB DATENNUTZUNG |
| | DATENVERARBEITUNG | NEB DATENNUTZUNG |

Anhaltspunkte für das Vorliegen einer Nebenbegriffsbeziehung sind:

- Nebenbegriffe haben einen gemeinsamen Oberbegriff:
  DATENSPEICHERUNG, DATENUEBERMITTLUNG, DATENVERAENDERUNG, DATENLOESCHUNG sind zueinander Nebenbegriffe und haben den gemeinsamen Oberbegriff DATENVERARBEITUNG.

Nebenbegriffe kommen in Aufzählungen vor.
Beispiel aus der Anlage zu § 6 Abs. 1 Satz BDSG:
1. ... (Zugangskontrolle), 2. ... (Abgangskontrolle),
3. ... (Speicherkontrolle), 4. ... (Benutzerkontrolle),
5. ... (Zugriffskontrolle), 6. ... (Übermittlungskontrolle),
7. ... (Eingabekontrolle), 8. ... (Auftragskontrolle),
9. ... (Transportkontrolle), 10. ... (Organisationskontrolle).
(Oberbegriffe zu diesen Begriffen sind: DATENSICHERUNGSMASSNAME, DATENSCHUTZMASSNAHME, KONTROLLMASSNAHME.)

Die Beispiele belegen den Zusammenhang der Nebenbegriffsbeziehung mit der Oberbegriffsbeziehung. Prinzipiell lässt sich die Nebenbegriffsrelation aus den hierarchischen Relationen automatisch ableiten.

zu (d): Morphosyntaktische Relationen

Diese Relationen setzen syntaktisch oder morphologisch unterschiedliche Ausprägungen eines Begriffs zueinander in Beziehung. Damit wird der Benutzer von der im Text verwendeten Variante unabhängig.

- Derivationsrelationen

Definition: Die Derivationsrelationen verknüpfen Wörter miteinander, die morphologisch voneinander oder von einem gemeinsamen dritten Wort ableitbar sind.
Als Relationenkennzeichen wird ein Kürzel aus 3 Buchstaben verwendet (D = Derivation, S = Substantiv, A = Adjektiv, V = Verb). Der 2. Buchstabe gibt die Wortklasse des Ausgangsbegriffs, der 3. Buchstabe die Wortklasse des Relatums an. Die Derivationsrelationen sind invertierbar (dargestellt durch Vertauschung des 2. und 3. Buchstaben im Relationskürzel).

| Beispiele: | | | |
|---|---|---|---|
| | STRUKTUR | DSA | STRUKTURELL |
| | SPEICHERN | DVS | SPEICHERUNG |
| | AUTOMATISIEREN | DVA | AUTOMATISCH |
| | AUTOMATISIERUNG | DSS | AUTOMATION |

Da Derivationsrelationen auch in der Version CTX-I (ohne semantische Differenzierung) verwendet werden können, war ihnen bereits ein besonderer Abschnitt (Kap. II.3.2) gewidmet.

- Zerlegungsrelationen

Definition: Über die Zerlegungsrelationen werden Komposita, Feste Wendungen und mehrwortige Begriffe mit ihren Teilwörtern verknüpft.
Bei den Relationen KOM, MEW und FEW ist der Ausgangsbegriff jeweils das Kompositum, der mehrwortige Begriff oder die Feste Wendung. Das Relatum ist der jeweilige Teilbegriff. Die Relationen sind invertierbar. Die Kürzel für die inversen Relationen lauten entsprechend TKM, TMW, TFW.

| | | |
|---|---|---|
| Beispiele: | DATENSCHUTZ | KOM DATUM1 |
| | DATENSCHUTZ | KOM SCHUTZ |
| | SCHUTZ | TKM DATENSCHUTZ |
| | SCHUTZ VON DATEN | FEW DATEN |
| | DATEN | TFW SCHUTZ VON DATEN |

Für Retrievalzwecke werden die Relationen TKM, TMW, TFW zu einer einzigen Relation zusammengefasst. Diese Relation spielt eine wesentliche Rolle in der Anwenderpraxis, da die Technik der "Truncation" hierdurch zuverlässig abgelöst werden kann. Voraussetzung ist dabei allerdings die Pflege der Zerlegungsrelationen.

- Abkürzungsrelationen

Definition: In der Relation ABK (Abkürzung) ist der Ausgangsbegriff eine Abkürzung (Kurzform) für das Relatum (Langform). Die Relation ist invertierbar (Invertierung: LNF (Langform)).

| | | |
|---|---|---|
| Beispiele: | BDSG | ABK BUNDESDATENSCHUTZGESETZ |
| | BILDSCHIRMTEXT | LNF BTX |

Diese Relation kann auch als eine spezielle Synonymierelation aufgefasst werden.

zu (e): Fachspezifische Relationen

DIN 1463 sieht ausdrücklich die Einbeziehung fachgebietsspezifischer Relationen vor. In der Laboranwendung Datenschutzrecht wurde beispielhaft eine entsprechende Relation implementiert.

- Relation Regelung - Regelungsgegenstand

Definition: In der Relation IUS bezeichnet der Ausgangsbegriff die juristische Regelung und das Relatum den zugehörigen Regelungsgegenstand. Die Relation ist invertierbar (Kürzel: REG). Dabei bezeichnet der Regelungsgegenstand i.a. einen Vorgang oder eine Tätigkeit, während die juristische Regelung eine Aussage darüber macht, ob dieser Vorgang durchgeführt werden soll, muss, darf oder ob ein Recht oder eine Pflicht zur Durchführung besteht.

| | | |
|---|---|---|
| Beispiele: | AUSKUNFT | REG RECHT AUF AUSKUNFT |
| | AUSKUNFTPFLICHT | IUS AUSKUNFT |

Hier sollen noch einige Erläuterungen zur Abgrenzung der Regelung-Regelungsgegenstand-Relationen (IUS und REG) von einer möglichen Tatbestand-Rechtsfolge-Relation (sie ist nicht im CTX-Thesaurus enthalten) gegeben werden.

Ein Beispiel soll den Unterschied erläutern (BDSG b 27, Abs. 1): "Personenbezogene Daten sind zu berichtigen, wenn sie unrichtig sind."

Der Tatbestand UNRICHTIGKEIT PERSONENBEZOGENER DATEN und die Rechtsfolge

PFLICHT ZUR BERICHTIGUNG PERSONENBEZOGENER DATEN wäre in eine "Tatbestand-Rechtsfolge-Relation" aufzunehmen, während durch die Relation IUS (bzw. invers REG) BERICHTIGUNG PERSONENBEZOGENER DATEN und PFLICHT ZUR BERICHTIGUNG PERSONENBEZOGENER DATEN zu verknüpfen sind (vorausgesetzt, diese Ausdrücke sind als mehrwortige Begriffe zur Relationierung zugelassen).

Eine Relation "Tatbestand-Rechtsfolge" wurde in die Laboranwendung des CTX-Systems zum Datenschutz bislang nicht aufgenommen, da sie erheblichen (gegenwärtig im Projekt nicht einbringbaren) intellektuellen Aufwand für Erstellung und Pflege erfordert hätte. Zudem erforderte sie in gewissem Umfang eine Gesetzesauslegung durch den Bearbeiter, die juristisch problematisch sein kann.

II. 3.6.2 Umfang

| | | | |
|---|---|---|---|
| Gesamtanzahl der relationierten Paare | | 48.568 | 100,00% |
| ohne Generierung | | 21.942 | 43,33% |
| mit Generierung | | 27.526 | 56,67% |
| davon durch: | Synonymie 51 ( 0,19%) | | |
| | Invertierung 27.475 (99,81%) | | |

Verteilung auf die einzelnen Relationentypen:

| | | |
|---|---|---|
| Gesamtzahl der relationierten Paare (ohne Invertierung) | 21.042 | 100,00% |
| Äquivalenzrelationen | 3.445 | 16,37% |
| Hierarchierelationen | 797 | 3,79% |
| Assoziationsrelationen | 3.624 | 17,22% |
| Morphosyntaktische Relationen | 13.097 | 62,24% |
| Fachspezifische Relationen | 79 | 0,38% |

Verteilung auf die einzelnen Relationen:

| | | |
|---|---|---|
| (a) Äquivalenzrelation | 3.445 | 100,00% |
| Synonymie | 402 | 11,67% |
| Mehrwortrelation | 988 | 28,68% |
| orthografische Varianten | 619 | 17,97% |
| Quasisynonymie | 1.165 | 33,82% |
| Übersetzungsrelation | 271 | 7,86% |
| (b) Hierarchierelation | 797 | 100,00% |
| Abstraktionsrelation | 572 | 71,77% |
| partitive Relation | 225 | 28,23% |
| (c) Assoziationsrelation | 3.624 | 100,00% |

| | | | |
|---|---|---|---|
| Assoziationsrelation | | 2.633 | 72,65% |
| Antonymrelation | | 456 | 12,58% |
| Kohyponymrelation | | 535 | 14,77% |
| (d) Morpho-syntaktische Relationen | | 13.097 | 100,00% |
| Derivationsrelationen: | | 1.878 | 14,34% |
| davon: | | | |
| Substantiv - Verb | 1.038 (7,93%) | | |
| Substantiv - Adjektiv | 568 (4,34%) | | |
| Verb - Adjektiv | 229 (1,75%) | | |
| Adjektiv - Adjektiv | 6 (0,05%) | | |
| Substantiv - Substantiv | 37 (0,27%) | | |
| Zerlegungsrelationen | | 10.965 | 83,72% |
| davon: | | | |
| Kompositum - Teil | 8.450 (64,52%) | | |
| Feste Wendung - Teil | 1.442 (11,01%) | | |
| Mehrwortiger Begriff - Teil | 1.073 ( 8,19%) | | |
| Abkürzungsrelation | 254 ( 1,94%) | | |
| (e) Fachspezifische Relationen | | 79 | 100,00% |

(Stand: Februar 1982)

II. 3.6.3 Kodierung

Da 1977 zu Beginn des Forschungsprojekts JUDO ein Thesaurus für das damals junge Fachgebiet "Datenschutz" nicht existierte, musste im Projekt mit der Konzipierung und Erstellung von semantischen Relationen neu begonnen werden. Anregungen zur Auswahl der Relationentypen und zur Vorgehensweise wurden den "Richtlinien für die Erstellung und Weiterentwicklung von Thesauri" sowie Werken wie HUTCHINS 1975 und LYONS 1972 entnommen. Aufgrund der Besonderheiten des Systems werden die herkömmlichen Relationen um Relationen aus dem morphosyntaktischen Bereich ergänzt. Ausgangsbasis zur semantischen Relationierung war eine Liste der dem Spezialgebiet "Datenschutz" zurechenbaren deskriptorfähigen Begriffe, die hauptsächlich der Loseblattsammlung BURHENNE/PERBAND 1970 entnommen waren.

Die Erweiterungsarbeiten am Thesaurus des Projekts JUDO-DS bauten auf diesen (konzeptionellen) Vorarbeiten auf. Bekanntlich handelt es sich bei der intellektuellen Erstellung und Pflege eines Thesaurus um eine sehr zeitaufwendige Arbeit, die zudem in weiten Bereichen umfangreiches Fachwissen erfordert. Ziel war es deshalb, Möglichkeiten zur Automationsunterstützung aufzudecken und arbeitssparende Verfahren zur Thesauruserstellung und Konsistenzprüfung zu entwickeln. Eine solche Automationsunterstützung konnte in verschiedenen Bereichen erreicht werden.

(a) Generierung von Relationen aufgrund formaler Eigenschaften

Einige Relationen können aufgrund ihrer formalen Eigenschaften automatisch erweitert werden.

Bei den symmetrischen Relationen (z.B. Synonymie) steht für jedes Wortpaar dieser Relation auch das invertierte Wortpaar in derselben Relation.

Bei inversen Relationen (z.B. Oberbegriff/Unterbegriff) steht für jedes Wortpaar der Relation R das invertierte Wortpaar in der dazu inversen Relation R'. Bei diesen Relationen muss nur eine Richtung intellektuell erstellt werden, die Umkehrung kann automatisch generiert werden. Da die verwendeten Retrievalsysteme TELDOK und GOLEM eine derartige Invertierung nicht unterstützen, müssen die inversen Relationen, obwohl sie implizit ineinander enthalten sind, explizit angegeben werden.

Eine Relation R heißt transitiv, wenn für alle Wortpaare (T1,T2) und (T2,T3), die in der Beziehung R stehen, gilt, dass auch die Begriffe T1 und T3 in der Beziehung R stehen. Transitivität kann man z.B. für die Synonymierelation oder die Abstraktionsrelationen annehmen (ist T1 Unterbegriff von T2 und T2 Unterbegriff von T3, dann kann man schließen, dass auch T1 Unterbegriff von T3 ist). Die Qualität dieser automatischen Erweiterungen hängt bei den "klassischen" Thesaurusrelationen zu einem beträchtlichen Teil von der "Schärfe" der Ausgangsrelationen ab. Schon bei intellektuell erstellten Relationen ist nicht immer Einigkeit über das Vorliegen einer behaupteten Beziehung zu erzielen; um so mehr können sich bei fortgesetzten automatischen Generierungen ursprünglich geringfügige Bedeutungsabweichungen zu unsinnigen Beziehungen summieren. Problemlos lässt sich die transitive Erweiterung jedoch z.B. bei den Derivationsrelationen durchführen.

(b) Erstellung von Derivations- und Zerlegungsrelationen

Die Zuordnung zwischen morphologischen Derivaten sowie zwischen Komposita und ihren Teilelementen wird von dem Zerlegungsalgorithmus des Analysesystems unterstützt. Der Zerlegungsalgorithmus liefert Zerlegungsvorschläge, die automatisch ausgewertet werden. Diese müssen anschließend intellektuell überprüft und insbesondere um die Bedeutungsdifferenzierung der Teilelemente ergänzt werden.

Insgesamt wird mit der Thesauruserstellung und -anwendung eine Effektivierung und Ökonomisierung der Dokumentdeskribierung durch Verlagern der intellektuellen Arbeit von der dokumentweisen Behandlung auf die Lexikonarbeit intendiert. Muss man bei der dokumentweisen Deskribierung bei zunehmender Dokumentmenge mit einem gleichbleibenden Arbeitsaufwand rechnen, so ist bei der Verlagerung der intellektuellen Arbeit auf die Wörterbucherstellung mit abnehmendem Arbeitsaufwand zu rechnen, da pro neuem Dokument nur die "neuen" Wörter bearbeitet werden müssen. (Man kann davon ausgehen, dass sich innerhalb eines Fachgebiets der Wortschatz, wenn auch erst bei großen Dokumentmengen - bis zu einem gewissen Grad sättigt).

## II.4 Das Programmpaket zur Deskriptorermittlung

### II.4.1 Überblick über das Verfahren

Aufgabe des Programmpaketes zur Deskriptorermittlung ist es, eine allgemeine Schnittstelle zwischen den Ergebnissen der linguistischen Textanalyse - ergänzt um zusätzliche (z.B. lexikalische) Informationen - und einem Information-Retrieval-System zu schaffen.

Die Ergebnisdaten, die von der linguistischen Analyse geliefert werden, lassen sich dazu gegen-

wärtig nicht unmittelbar (d.h. ohne weitere Aufbereitung) verwenden. Die zur Dokumentdeskribierung benötigten Informationen, z.B. eine Auswahl der syntaktischen Relationen, werden daher über Algorithmen aus der bei der Analyse aufgebauten Datenstruktur auf eine allgemeine Input-Schnittstelle für ein Retrieval-System abgebildet (umgesetzt). Die automatische linguistische Analyse liefert als Bestandteil eines automatischen Übersetzungssystems zum Teil Informationen, die zwar bei einem Transfer von der Quell- in die Zielsprache erforderlich sind, aber für die Deskribierung von Dokumenten derzeit nicht ausgewertet werden. Dazu gehören wortspezifische Merkmale wie Angaben über Genus, Kasus, Numerus, Valenzen, sowie satzstrukturspezifische Angaben wie Aktiv-/Passiv-Transformation, Neben-, Infinitiv- und Relativsatzanschlüsse usw. Um die Verarbeitungsgeschwindigkeit der Analyse zu erhöhen, werden zudem alle alphanumerischen Zeichen in einen speziellen Code umgewandelt und in einem besonderen, von dem verwendeten Rechner abhängigen Datenformat ausgegeben.

Bei der Überführung in die allgemeine Datenschnittstelle werden zur Zeit nicht verwertbare Informationen ausgefiltert, die internen Codes werden auf eine allgemeine, durch Programme in höheren Programmiersprachen unmittelbar weiterverarbeitbare Form umgesetzt. Von der allgemeinen Schnittstelle (Zwischenformat) aus werden die Daten durch spezifische Anpassungsprogramme (vgl. Kap. II.6) für das jeweilige IR-System (z.Z. GOLEM und TELDOK) in die jeweils benötigte Form umgesetzt.

Im Rahmen der Systementwicklung wurden an die Deskriptorerstellung folgende Anforderungen gestellt:

- Die Darstellung (Outputschnittstelle) sollte unabhängig von einem konkret zu verwendenden Information-Retrieval-System sein.

- Informationen, die für verschiedene Information-Retrieval-Systeme notwendig sind, sollten nur einmal erstellt werden.

- Eine interaktive Korrektur- bzw. Selektionskomponente sollte bei diesem Zwischenformat einsetzen und nicht auf spezielle IR-Systeme abgestimmt werden müssen; einmal durchgeführte Korrekturen sollten für alle IR-System-Versionen gelten.

Das Zwischenformat speichert als relativ "universale" Schnittstelle alle Zusatzinformationen, die in einer der heute am Markt üblichen Retrieval-Versionen von Nutzen sein können. Die Informationen gehen also im Regelfall über diejenigen hinaus, die von einem bestimmten IR-System ausgewertet werden können (oder sollen: dies ist oft auch eine Akzeptanzfrage und/oder Frage der Benutzerschulung/Benutzerkenntnisse).

Eingabedatenfür die Deskriptorerstellung sind

- Ergebnisse der automatischen Sprachanalyse;
- Fachlich-lexikalische Informationen (z.B. fachspezifische Informationen; Definitionen von Bedeutungsvarianten);
- Zuordnungsdaten (Zuordnung von Satznummern eines Textes zu den jeweiligen Dokumentnummern zur Identifikation von Dokumenteinheiten).

Ausgabedaten der Deskriptorermittlung sind

- Einfache Deskriptoren (Wortlaut mit deskriptor- und textspezifischen Zusatzinformationen);
- Komplexe Deskriptoren (Einfache Deskriptoren in syntaktischen Relationen);
- Kontrolldaten (Zwischenergebnisse, Statistiken und Fehlermeldungen).

### II.4.2 Einfache Deskriptoren

Einfache Deskriptoren umfassen die Deskriptorwortlaute mit Identifikations-Informationen sowie linguistische, fachlich-lexikalische und technische Zusatzinformationen.

#### II.4.2.1 Deskriptorwortlaut

Deskriptorwortlautist die lemmatisierte Form, d.h. die Grundform des Textwortes, wie sie sich aus dem Sprachanalyseverfahren ergibt. Hierzu gehören auch Deskriptorwortlaute, die bei der semantischen Analyse durch die Zusammenführung von Einzelwörtern zu Festen Wendungen (z.B. JURISTISCH und PERSON zu JURISTISCHE PERSON) gebildet werden.

Beispiele:

| | |
|---|---|
| DATUM | (Formaldeskriptor) |
| DATUMI | (bedeutungsdifferenzierter Einfacher Deskriptor) |
| JURISTISCH | ("normaler" Einfacher Deskriptor) |
| JURISTISCHE PERSON | ("Feste Wendung" als Einfacher Deskriptor |
| DATENSCHUTZ | (Kompositum als Einfacher Deskriptor) |

#### II.4.2.2 Identifikationsinformationen

Zu den Identifikationsinformationen gehören:

- Textkennzeichnung
- Nummer des Dokuments im Text
- laufende Satznummer im Text
- laufende Satznummer im Dokument
- Wortnummer im Satz mit Unterscheidungskennung

Die Textkennzeichnung dient zur formalen Unterscheidung verschiedener Texte. Derartige Kennzeichnungen können bei der Umsetzung in eine Datenbank verwertet werden. Sie ermöglichen aber auch, einzelne Texte hinsichtlich gemeinsamer oder unterschiedlicher Deskriptoren zu vergleichen. Damit können auf der Deskriptorebene die Entwicklungen eines Gesetzes vom Entwurf bis zur verabschiedeten Fassung oder die Gesetze einzelner Bundesländer einfach miteinander verglichen werden.

Die Nummer des Dokuments wird für jeden einzelnen Text aufgrund einer automatisch erstellten Zuordnungstabelle aus den laufenden Satznummern im Text vergeben. Zur Einteilung in Dokumente bietet sich bei Gesetzen der einzelne Paragraph an; bei anderen Texten (z.B. Berichten, Kommentaren) erfolgt die Einteilung nach inhaltlich-informativen Zusammenhängen. Gegen-

wärtig wird darauf geachtet, dass bei Berichten und Kommentaren (also soweit nicht schon "natürliche" Einheiten vorliegen, wie bei "Paragraphen" oder "Lexikonstichwort") die einzelnen Dokumente weder zu "groß" bzw. zu "klein" werden, da sonst die Gefahr besteht, bestimmte Dokumente zu häufig (wegen der zu großen Anzahl von Deskriptoren) bzw. zu selten beim Retrieval mit verknüpften Deskriptoren zu treffen. Letztlich ist die Abgrenzung von Dokumenten eine Aufgabe der Benutzerforschung und damit systemunabhängig. Es wäre sogar denkbar, dass eine Dokumenteinteilung erst bei der Umsetzung in eine konkrete Informationsbank erfolgt oder sogar im Information-Retrieval-System selbst variabel gestaltet wird.

Die laufende Satznummer im Text sowie die Wortnummer im Satz ergibt sich aus der Wort- und Satzsegmentierung des Satzanalysebausteins zur Texteingabe; die laufende Satznummer im Dokument wird zusätzlich aus der jeweiligen laufenden Satznummer im Text ermittelt. Die Vergabe der Satz- und Wortnummer ermöglicht es prinzipiell (unabhängig von den syntaktischen Analyse-Ergebnissen), nach formalen Kriterien zu recherchieren wie "Deskriptor kommt im gleichen Satz vor", "Deskriptor ist von einem anderen Deskriptor n Worte entfernt", also unter Verwendung entsprechender Funktionen (z.B. Adjacent-Angabe) wie bei STAIRS oder DIRSGRIPS, bzw. entsprechende Statistiken aufzubauen.

Die Unterscheidungskennung bei einer Wortnummer dient dazu, Festen Wendungen sowie Bedeutungsvarianten eines Wortes die gleiche Wortnummer wie dem Ausgangswort zuzuordnen.

### II.4.2.3 Linguistische Informationen

Diese Informationen umfassen:

- Wortklasse des Stammwortes
- aktuelle Wortklasse im Text
- Kennzeichen für attributives Partizip
- Nummer des Subsatzes
- Nummer der Verbal- bzw. Nominalgruppe
- Bedeutungsnummer des Deskriptors
- Analysetiefe der Satzanalyse

Durch die morpho-syntaktische Analyse werden die Wortklassen eines Textwortes bestimmt; es wird dabei unterschieden zwischen der Wortklasse des Stammwortes (z.B. Verb) und der aktuellen Wortklasse im Text (z.B. Substantivierter Infinitiv, Partizip Perfekt; vgl. hierzu auch SATAN-Handbuch, Kapitel B). Ein Textwort ist deskriptorrelevant, wenn es einer der Stammwortklassen Adjektiv (incl. Zahlwort), Verb oder Substantiv angehört. Es werden also alle Funktionswörter wie Artikel, Präpositionen, Pronomina, Hilfsverben etc. als Deskriptorkandidaten verworfen. (Man nützt damit die Wortklassenkennung als eine Art Stoppwort-markierung aus.)

Ein besonderes Problem ergibt sich bei attributiv gebrauchten Partizipien, also Wörtern der "Stammwortklasse" Verb und der aktuellen (d.h. an der Satzoberfläche auftretenden) Textwortklasse Adjektiv. Durch die Analyse werden diese Wörter, unabhängig davon, ob es sich um Partizip I (Präsens) oder Partizip II (Perfekt) handelt, immer auf den Infinitivstamm zurückgeführt, so dass sich sowohl für die Phrase "speichernden Stellen" als auch für die Phrase "gespeicherte Stelle" als Deskriptoren SPEICHERN und STELLE ergeben. Durch die Analyse werden jedoch

Informationen über die Art des attributiven Partizips gewonnen. Damit ist es möglich, zusätzlich (neben dem Verb) die entsprechende Partizipform, also SPEICHERND bzw. GESPEICHERT zu erzeugen; diese Partizipform wird dann zur Generierung des entsprechenden Komplexen Deskriptors (hier: SPEICHERNDE STELLE bzw. GESPEICHERTE STELLE) verwendet.

Die Wortklasseninformation als solche wird von den im Rahmen des JUDO-Projekts realisierten Information-Retrieval-Systemen nur in der GOLEM-Version verwendet; darüber hinaus dient sie gegenwärtig vorwiegend statistischen Zwecken.

Die Zugehörigkeit eines Wortes zu einem (auch diskontinuierlichen) Subsatz sowie die Verbal- und Nominalgruppen werden durch die syntaktische Analyse ermittelt. Diese Informationen können z.B. bei GOLEM in der Feinrecherche (s.o.) zur Präzisierung von Retrievalanfragen verwendet werden.

Die Bedeutungsnummer eines Deskriptors, die durch die semantische Analyse bestimmt wird, ermöglicht die Unterscheidung von Bedeutungsvarianten eines Homonyms. So kann z.B. ANLAGE die verschiedenen Bedeutungen im Sinne von "Rechenanlage", "Schriftzusatz", "Parkanlage", "Geldanlage, "Gebäude", "Talent" usw. haben, während dem indizierten Deskriptor ANLAGE1 (= ANLAGE mit Bedeutungsnummer 1) nur noch die Bedeutung "Rechenanlage" oder "Computer" zugeordnet ist.

Die Analysetiefe gibt an, welches der zuletzt erfolgreich durchlaufene Analyse-Programmteil war. Die Vereindeutigung eines Wortes, die Bildung einer Festen Wendung oder einer syntaktischen Relation ist nur möglich, wenn ein Satz zumindest noch durch den Programmteil zur Ermittlung der Nominalgruppen bearbeitet wurde. Dies erleichtert die qualitativen Kontrollen der Analyseergebnisse.

#### II.4.2.4 Fachlich-lexikalische Informationen

Diese Informationen wurden zuvor durch den "Fachexperten" (in der vorliegenden Testanwendung von CTX also durch einen Juristen) intellektuell vorgegeben (vgl. Kap. II.3.5). Es handelt sich im Einzelnen um die Ergänzung von Wörtern durch Zusatzinformationen über

- Fachgebietszuordnung
- Relevanzkennzeichnung
- fachtextbezogene Gewichtungskennzeichnung (Wahrscheinlichkeitsangabe)

In einer ersten Systemrealisierung wurden auch die Ergebnisse der Kompositumzerlegung in sinnvolle, relevante Einzeldeskriptoren mit einer Zerlegungskennzeichnung direkt dem einzelnen Dokument zugeordnet. Dies hatte jedoch zur Folge, dass gerade bei mehrgliedrigen Komposita besonders viele Deskriptoren aus Zerlegungsergebnissen einem Dokument zugeordnet wurden. Bei einer evtl. fehlerhaften Zerlegung eines Kompositums ist zudem eine spätere Korrektur auf Dokumentebene erschwert. Die Kompositumzerlegung wurde daher - wie bereits erwähnt - in Form einer Relation in den Thesaurus des entsprechenden Datenbanksystems integriert. Diese Zuordnung von Zerlegungsergebnissen mittels einer Relation konnte jedoch nur bei der GOLEM-Version (vgl. Kap. II.6.1) realisiert werden, da bei der TELDOK-Version die zur Verfügung stehenden Relationen (maximal 6) bereits voll genutzt waren.

### II.4.2.5 Technische Informationen

Hierunter sind Informationen zusammengefasst, die nicht in die weitere Datenbankverarbeitung eingehen, sondern Hilfsmittel hei der technischen Verarbeitung darstellen; hierzu zählen auch Daten, die für Statistikzwecke herangezogen werden. Dazu gehören:

- Angaben zur Zeichenlänge eines Deskriptors (bei mehrwortigen Deskriptoren (Festen Wendungen) einschließlich der Leerzeichen zwischen den Wörtern),

- Anzahl der syntaktischen Relationen, in denen dieser Deskriptor linkes und/oder rechtes Element ist,

- Kennzeichnung, ob dieser Deskriptor (mit seinen Zusatzinformationen) intellektuell korrigiert/erweitert wurde (gegenwärtig wird eine intellektuelle Korrektur noch nicht durchgeführt).

## II.4.3 Auswertung syntaktischer Strukturen zur Dokumentbeschreibung (Komplexe Deskriptoren)

Neben den oben beschriebenen sog. Einfachen Deskriptoren werden weitere Deskriptoren gebildet, für deren Erstellung die von der Sprachanalyse gewonnenen syntaktischen Oberflächenstrukturen ausgewertet werden. Ausgewählt wurden hierbei (vorerst) Beziehungen zwischen

- Adjektiv-Attribut und einem Nomen (ADJ-Relation)
- Nomen und Genitiv-Attribut (GEN-Relation)
- Nomen und Präpositional-Attribut (PRP-Relation)
- Nomen und nominaler Anreihung (KON-Relation)
- Verb und Akkusativkomplement (VRB-AKK-Relation)
- Verb und Modalverb (VRB-MOD-Relation)

Komplexe Deskriptoren werden gebildet, indem zwischen die beteiligten Einfachen Deskriptoren als Relator ein charakterisierender Buchstabe (G, P, K entsprechend für GEN-, PRP-, KON-Relation) als Abkürzung der syntaktischen Relation gesetzt wird. Weitere Möglichkeiten sind die Rekonstruktion der Oberflächenform (ADJ-Relation) bzw. das Nebeneinanderstellen der beteiligten Einfach-Deskriptoren (VRB-AKK-Relation, MOD-VRB-Relation).

Hinzugefügt werden die in Kap. II.4.2.2 beschriebenen Identifikationsinformationen für die Zuordnung zwischen Text, Dokument und Satz.

### II.4.3.1 ADJ-Relation

Beim Aufbau der Beziehung zwischen einem Adjektiv(-Attribut) und einem Nomen wird nicht nur das unmittelbar vor dem Nomen stehende Adjektiv verarbeitet, sondern sämtliche zu diesem Nomen gehörenden Adjektive, die durch die Analyse als nebengeordnet gekennzeichnet sind.

Beispielsweise liefert die in § 2 BDSG vorkommende Sequenz

"(...) gespeicherter oder durch Datenverarbeitung unmittelbar gewonnener Daten (...)"

die folgenden Komplexen Deskriptoren:

GESPEICHERTES DATUM
GEWONNENES DATUM

II.4.3.2 GEN-Relation

Hierbei handelt sich um eine Beziehung zwischen einem Nomen und seinem zugehörigen Genitiv-Attribut.

Beispielsweise ergibt die Phrase "Unternehmen der Presse" den Komplexen Deskriptor UNTERNEHMEN G PRESSE .

Die GEN-Relation wird nicht nur in den Fällen gebildet, in denen die bei der Relation beteiligten Nomina nur durch einen Artikel getrennt sind, sondern auch bei Textfolgen mit mehrwortigen Einschüben zwischen den beteiligten Nomina; als Beispiel kann hier folgender Textausschnitt (aus § 2 BDSG) angeführt werden:

"Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person"

Hier wird (neben den ADJ-Relationen bei Person) der Komplexe Deskriptor VERHAELTNIS G PERSON ermittelt.

II.4.3.3 PRP-Relation

Bei dieser Relation ist statt eines Genitiv- ein Präpositionalattribut beteiligt.

Beispielsweise ergibt die Textfolge "Einzelangaben über Verhältnisse" den Komplexen Deskriptor EINZELANGABE P VERHAELTNIS.

In einem Beispiel (§ 40 BDSG) mit längerem Einschub: "Register über die nach § 39 Absatz 1 anmeldepflichtigen Stellen" wird (neben anderen Relationen) die Relation REGISTER P STELLE aufgebaut.

II.4.3.4 KON-Relation

Hierbei handelt es sich um eine Beziehung zwischen einem Nomen und seiner nominalen Anreihung. Diese Relation ist im Gegensatz zur GEN- und PRP-Relation symmetrisch und transitiv. Symmetrisch heißt dabei, dass aus "NOM1 K NOM2" auch "NOM2 K NOM1" folgt; transitiv heißt, dass aus "NOM1 K NOM2" und "NOM2 K NOM3" auch "NOM1 K NOM3" folgt.

Von der linguistischen Analyse werden bei einer nominalen Anreihung von links nach rechts die syntaktischen Beziehungen nur zwischen zwei aufeinander folgenden Nomina hergestellt. Die Erzeugung der symmetrischen und transitiven KON-Relationen erfolgt durch einen einfachen Algorithmus.

Als Textbeispiel (§ 1 BDSG) sei hier eine nominale Anreihung mit 4 Elementen angeführt:

"Speicherung, Übermittlung, Veränderung und Löschung" ergibt die Komplexen Deskriptoren

| | | |
|---|---|---|
| SPEICHERUNG | K UEBERMITTLUNG | (einfach) |
| UEBERMITTLUNG | K VERAENDERUNG | ( " ) |
| VERAENDERUNG | K LOESCHUNG | ( " ) |
| SPEICHERUNG | K VERAENDERUNG | (transitiv) |
| SPEICHERUNG | K LOESCHUNG | ( " ) |
| UEBERMITTLUNG | K LOESCHUNG | ( " ) |
| UEBERMITTLUNG | K SPEICHERUNG | (symmetrisch) |
| VERAENDERUNG | K UEBERMITTLUNG | ( " ) |
| LOESCHUNG | K VERAENDERUNG | ( " ) |
| VERAENDERUNG | K SPEICHERUNG | ( " ) |
| LOESCHUNG | K SPEICHERUNG | ( " ) |
| LOESCHUNG | K UEBERMITTLUNG | ( " ) |

Damit ergeben sich aus 4 Nomina bzw. Einfachen Deskriptoren noch 6 weitere Komplexe Deskriptoren. Auf die Erzeugung der transitiven KON-Relation kann nicht verzichtet werden, da die Reihenfolge von angereihten Nomina häufig eher zufällig ist und sogar innerhalb eines Textes variiert. Dies gilt dementsprechend auch für eine Suchanfrage. Solange mit herkömmlichen Information-Retrieval-Verfahren gearbeitet wird, also beim Vergleich einer Suchfrage mit Deskriptoren operiert werden muss und nicht etwa - was prinzipiell denkbar wäre (vgl. z.B. das CONDOR-System) - wiederum komplexe Algorithmen zur Verfügung stehen, die derartige Symmetrie- und Transitivitätsrelationen im konkreten Fall berücksichtigen, erscheint diese explizite Umformung bzw. Erweiterung (auf Kosten des Speicherbedarfs) bei dieser Schnittstelle erforderlich. Auf die invertierten komplexen Deskriptoren kann verzichtet werden, wenn diese bei der Integration in ein herkömmliches Information-Retrieval-System als "Synonyme" erfasst werden. Systeme, die diese Invertierung bei der Recherche algorithmisch durch Integration einer entsprechenden Prozedur in den Retrieval-Vorgang lösen, sind bislang für die Praxis noch nicht verfügbar.

### II.4.3.5 Erweiterung der nominalen Relationen

Wichtig erscheint die Vervollständigung der vorhandenen Relationen durch Generierung "getilgter" Elemente (im linguistischen Sinne), z.B. bei nominalen Anreihungen. Dazu sind Regeln zu erarbeiten, die es etwa erlauben, dass bei solchen Anreihungen das getilgte Adjektiv-, Genitiv- oder Präpositional-Attribut erkannt und dementsprechend an der Oberfläche generiert wird. So müssten in den folgenden Beispielen

"bestimmte Stellen oder Behörden",
"Unternehmen oder Hilfsunternehmen der Presse",
"Behörden des Bundes und der Länder",
"Daten und Informationen über Personen",
"Register über Firmen und Unternehmen"

die fehlenden Attribute wie folgt ergänzt werden:

"bestimmte Stellen oder bestimmte Behörden",
"Unternehmen der Presse oder Hilfsunternehmen der Presse",
"Behörden des Bundes und Behörden der Länder",
"Daten über Personen und Informationen über Personen",
"Register über Firmen und Register über Unternehmen."

Es gibt jedoch auch Beispiele dafür, dass bei nominalen Anreihungen Attribute nicht wiederholt werden dürfen, z.B. bei

"öffentliche Stellen und Kirchen",
"Verstöße gegen Vorschriften und Mängel",
"Kirchen und Behörden des Bundes".

In einem ersten Test wurden auf der Basis des BDSG bei nominalen Anreihungen automatisch alle formal möglichen Attribute ergänzt. Dabei traten unter anderem die eben angeführten Fehlertypen auf.

#### II.4.3.6 VRB-Relationen

Zusätzlich zu den nominalen Strukturbeschreibungen wurden testweise auch verbale Strukturen aufgebaut (Beispiel: Auskunft erteilen VRB-AKK-Relation, oder BERICHTIGEN DUERFEN - VRB-MOD-Relation).

Diese Relationen dienen der Vorbereitung für Algorithmen, die über die Integration von Thesaurusrelationen eine gemeinsame Abbildung der folgenden Strukturen ermöglichen sollen:

AUSKUNFTSERTEILUNG - AUSKUNFT ERTEILEN - ERTEILUNG VON AUSKUNFT
im Fall der VRB-AKK-Relation bzw.

BERICHTIGEN MUESSEN - BERICHTIGUNGSPFLICHT - VERPFLICHTUNG ZUR BERICHTIGUNG
im Fall der MOD-VRB-Relation (vgl. dazu auch Kap. I.2.6).

### II.5 Probleme des dokumentarischen Überbaus

Für die Laborimplementierung sollte eine möglichst realitätsnahe Form gefunden werden. Da der Bereich "Datenschutzrecht" zugrunde gelegt wurde, lag es nahe, die Datenbank an der JURIS-Implementierung zu orientieren. Diese schließt einen sog. "dokumentarischen Überbau" mit ein. Während CTX auf die Ermittlung von Freitext-Deskriptoren ausgerichtet ist, wird üblicherweise in gängigen IR-Systemen - so auch bei JURIS - eine umfangreiche Menge an strukturierten Informationen bereit gestellt, die v.a. aus sog. "aspektgebundenen" Deskriptoren besteht. Um eine möglichst realistische Benutzersituation zu simulieren - aber auch im Hinblick auf die möglichen Präzisierungen (z.B. über Dokumenttypen) für Retrievaltests - wurde auch für die Datenbank von JUDO-DS (Golem-Version) je Dokument ein derartiger Überbau erstellt.

Die Informationswiedergewinnung in der Laborimplementierung "Datenschutzrecht" des CTX-

Systems ist somit auf zweifache Weise möglich, nämlich über freie und aspektgebunde Deskriptoren. An dieser Stelle soll auf die aspektgebundenen Deskriptoren eingegangen werden, die im wesentlichen diesen dokumentarischen Überbau repräsentieren.

### II.5.1 Suchhilfen und Suchkriterien

Grundlage für eine erfolgreiche Informationswiedergewinnung sind zwei Voraussetzungen, einerseits Suchhilfen, andererseits Suchkriterien. Als Suchhilfen dienen die Richtlinien, die als formale Voraussetzungen bei der Aufbereitung und der Eingabe des Dokuments erarbeitet wurden, wie Schreibweise, Wortformen, Wortzusammensetzungen, Abkürzungen usw. Suchkriterien werden durch Sinn und Zweck des Informationssystems vorgegeben. Äußerer Sinn und Zweck der Modellanwendung von CTX im Bereich Datenschutz ist es, Dokumente für Juristen und andere daran Interessierte aufzubereiten. Dementsprechend müssen sich die Suchkriterien an die für den deutschen Rechtskreis geltenden Zitierweisen halten, wie z.B. Normen- oder Aktenzeichen. In anderen Rechtskreisen mögen andere Suchkriterien notwendig sein, wie es beispielsweise im angloamerikanischen ausreichend ist, die Namen der Prozessbeteiligten anzugeben, um ein Dokument ausreichend zu identifizieren. Diese juristischen Suchkriterien werden in der GOLEM-Anwendung formalisiert, indem sie als Deskriptor an einen Aspekt (wie z.B. Aktenzeichen oder Normenkette) gebunden werden.

### II.5.2 JURIS-aspektgebundene Deskriptoren

Der Überbau wurde in der GOLEM-Anwendung von CTX auf der Grundlage der von JURIS entwickelten Aspekte erstellt. Bei JURIS sind diese inzwischen auf die Anzahl von über 100 angewachsen, die sich aus Aspekten für Rechtsvorschriften, für Gerichtsentscheidungen und für Literatur zusammensetzen. Es war jedoch nicht notwendig, im Rahmen der Modellanwendung von CTX alle diese Differenzierungen nachzuvollziehen, da eine weit geringere Anzahl von Dokumenten unterschieden werden musste. So wurde als Anhalt der Überbau des Textteils von JURIS verwendet, der um einige Aspekte des Suchwortteils von JURIS erweitert wurde. Um Strategien zur Informationswiedergewinnung analog den Möglichkeiten von JURIS zu gewährleisten, wurden dabei v.a. jene JURIS-Aspekte ausgewählt, die mit den in der Rechtswissenschaft und Rechtspraxis üblichen Zitierweisen übereinstimmen. Die von JURIS verfolgte Entwicklung, die die herkömmlichen unterschiedlichen Dokumentationsweisen der verschiedenen Urheber und Verbreiter berücksichtigt, wurde nicht übernommen. Die verarbeiteten unterschiedlichen Textsorten (vgl. Kap. II.1.1) führen dabei zu einer unterschiedlichen Aspektierung.

### II.5.3 Art der Aspekt-Reduzierung

Die Reduzierung der Aspekte erfolgte einmal dadurch, dass nur die juristische Zitierweise aufgenommen wurde. Weiterhin wurde auf die Mehrfachvergabe eines Aspekts verzichtet, die bei JURIS durch das eigens entwickelte Programmsystem PARAT automatisch erzeugt wird.

Ebenfalls außer acht gelassen wurde der Aspekt "Schlagwort". Ziel des CTX-Systems ist es ja gerade, auch zu einer Reduzierung der intellektuellen Verschlagwortung beizutragen, indem morphosyntaktische Strukturen eingebracht und Thesaurusrelationen gebildet werden. Diese Entwicklung soll dazu führen, den intellektuellen Aufwand für die Verschlagwortung zu reduzieren bzw. entfallen zu lassen. Diese Möglichkeit ist insofern nahe liegend, als sich bei JURIS

zeigte, dass in der Regel die Zentralbegriffe der Langtexte vom Bearbeiter als Schlagworte vergeben werden. Diese Zentralbegriffe sind fast immer Komposita oder morphosyntaktische Strukturen, die zu einem Kompositum zusammengefasst werden können. Derartige Verschlagwortungen könnten somit durch das CTX-System zum Teil oder insgesamt automatisiert werden.

Weiterhin wurde auf Aspekte verzichtet, die nur für das JURIS-System selbst von Bedeutung sind.

### II.5.4 Beispiel für einen CTX-Überbau

Als Beispieltext für den Abschlussbericht wurden zwei Textstellen aus dem "Vierten Bericht über die Tätigkeit des Landesbeauftragten für den Datenschutz des Saarlandes" (Beispieltext) ausgewählt.

Aspektgebundene Deskriptoren zu Dokument 1:

Art: Literatur
Art: selbständige Literatur Datum: 1983-01-04
Gliederungstitel: 9.2 Datenschutzprobleme
Gliederungstitel: Datenschutzprobleme
Fundstelle: Landtagsdrucksache 8/1189
Fundstelle: Landtagsdrucksache
Haupttitel: Vierter Bericht über die Tätigkeit des Landesbeauftragten für den Datenschutz
Herausgeber: Landtag des Saarlandes
Herausgeber: Saarländischer Landtag
Titel: Tätigkeitsbericht
Titel: Datenschutz
Titel: Landesbeauftragter für Datenschutz
Typ: Tätigkeitsbericht
Verfasser: Der Landesbeauftragte für Datenschutz des Saarlandes
Verfasser: Saarländischer Landesbeauftragter für Datenschutz

Die aspektgebundenen Deskriptoren zu Dokument 2 sind dieselben wie zu Dokument 1 mit Ausnahme von:

Gliederungstitel: 9.4 Regelungsbedarf
Gliederungstitel: Regelungsbedarf

## II.6 Datenumsetzung für Information-Retrieval-Systeme

Ein Aufgabenbereich des Projekts JUDO-DS umfasste die Einbringung der Indexierungsergebnisse in (schon vorhandene) Retrievalsysteme. Dazu wird von einer allgemeinen Schnittstelle ausgegangen, in der alle vorhandenen Daten vorliegen. Sie ist damit Grundlage für die Umsetzung der Ergebnisse auf die Erfordernisse der verschiedenen konkreten Retrievalsysteme (vgl. Kap. II.4).

Da es Zielsetzung des Projekts war, keine neue Retrievalsoftware zu entwickeln, sondern am Markt befindliche Systeme zu benutzen - die Retrievalsysteme als Fremdsoftware also vorge-

geben sind - bestand die Aufgabe darin, die in der Schnittstelle vorliegenden Ergebnisse so auf die jeweilige Daten- bzw. Informationsbankstruktur abzubilden, dass dabei möglichst wenig Informationen verloren gehen bzw. möglichst viele der vorliegenden Informationen für ein Retrieval nutzbar gemacht werden können.

Die Realisierung der Indexierungsergebnisse in konkreten Retrievalsystemen (hier TELDOK und GOLEM) soll dazu dienen, praktische Tests zur Qualität der Indexierungsverfahren durchzuführen, die zugleich Aufschlüsse geben sollen über die Anforderungen, die ein spezifischer Benutzerkreis an ein IR-System stellt. Nicht verfolgt wurden in diesem Projekt die Anforderungen der Benutzer an die technische Seite des Retrievalsystems selbst (sie sind z.T. schon anderweitig festgehalten, vgl. z.B. PLESCH/GRIESE 1972); jedoch kann eine Auswertung der Tests auch auf dieser Ebene zu Anforderungen führen, die an ein modernes Retrievalsystem (z.B. hinsichtlich der Darstellbarkeit sprachlicher wie lexikalischer Strukturen) zu stellen sind (man denke an eine Recherchekomponente mit einer Schnittstelle zur natürlichsprachigen Problembeschreibung, vgl. z.B. CONDOR (WIELAND 1979}.

Der Schwerpunkt bei den noch ausstehenden praktischen Tests (wie sie z.B. im nachfolgenden Projekt TRANSIT für verschiedenste Anwendungsbereiche durchgeführt werden) wird darin liegen, Aufschlüsse über Qualität und Akzeptanz der angebotenen Indexierung zu erhalten. Zugleich sollen Einblicke gewonnen werden darüber, mit welchen Suchstrategien die Benutzer bei der Informationsgewinnung vorgehen. Demzufolge ist es erforderlich, bei der Abbildung der Indexierungsergebnisse auf die DB-Strukturen, d.h. dort, wo es die Retrievalsysteme zulassen, mehrere Möglichkeiten zu realisieren, somit also dem Benutzer erst einmal im Modell - gleichsam als Spielwiese - die Möglichkeit zu eröffnen, sich eine Suchstrategie im Rahmen des jeweiligen Retrievalsystems auszuwählen.

Ein Beispiel dazu ist die Darstellung syntaktischer Informationen einmal in Form Komplexer Deskriptoren und in Form von Deskriptoren-Indizes. Beide Darstellungen ermöglichen die Verfolgung unterschiedlicher Suchstrategien; das eine Mal auf der Ebene der Grobrecherche, das andere Mal auf der Ebene der Feinrecherche (eine andere Variante - die unmittelbare Integration von Problembeschreibungen in natürlicher Sprache in den Retrievalprozess - ist im Verfahren CONDOR realisiert (vgl. BANERJEE 1977). Ansätze dazu sind bereits in dem Deskriptorermittlungsverfahren aus natürlichsprachiger Problembeschreibung (NATURA) vorhanden, das jedoch noch nicht unmittelbar in das Retrieval integriert ist.

### II.6.1 Umsetzung GOLEM

#### II.6.1.1 Technische Grundlagen

Im Rahmen des Projekts JUDO wurde als eine konkrete Retrievalkomponente eine Informationsbank mithilfe des Systems GOLEM eingerichtet (JUDO-G).

Das GOLEM-System war zunächst auf der Siemens-Rechenanlage im Rechenzentrum des Bundesministeriums der Justiz in Bonn verfügbar; der Anschluss daran erfolgte über eine Standleitung, die seit Projektbeginn (1977) zwischen dem Rechenzentrum in Köln (jetzt in Bonn) und der Abteilung für Nichtnumerische Datenverarbeitung an der Universität Regensburg als wissenschaftlichem Anwender von JURIS eingerichtet war. In Erweiterung der sonst üblichen Teilhabersystemanschlüsse bestand bei diesem Terminalanschluss die Möglichkeit, Dateien auf der Re-

chenanlage einzurichten, eine wesentliche Voraussetzung, um eine Informationsbank aufzubauen, und dazu nötige Programme zu implementieren. Nach der Migration des Projekts nach Saarbrücken konnte die GOLEM-Implementierung auf dem Rechner der Universität des Saarlandes genutzt werden.

#### II.6.1.2 Möglichkeiten von GOLEM im Hinblick auf eine Anwendung der Indexierungsergebnisse von CTX

Ausgangspunkt ist eine Schnittstelle (vgl. Kap. II.4), in der die Indexierungsergebnisse weitestgehend maschinenunabhängig in einer allgemeinen Form, d.h. nicht spezialisiert für eine Datenbankanwendung, vorliegen.

Die Indexierungsergebnisse beinhalten:

(1) den Inhalt des Textes repräsentierende Deskriptoren auf ihre Grundform lemmatisiert,
(2) semantische Informationen,
(3) syntaktische Informationen.

In GOLEM können die Begriffe praktisch ohne Längenbeschränkung auf GOLEM-Deskriptoren abgebildet werden (im Unterschied zu TELDOK, das eine Längenbeschränkung für Deskriptoren auf 29 Zeichen vorschreibt). Das bedeutet, dass auch Deskriptoren gebildet werden können, die aus mehreren Wörtern bestehen, ohne dass man (wie bei der TELDOK-Anwendung) gezwungen ist, diese abzukürzen. Dies ist insbesondere relevant für die Bildung Komplexer Deskriptoren und die Darstellbarkeit Fester Wendungen als Deskriptoren.

Die vorliegenden semantischen Informationen - Disambiguierung semantisch mehrdeutiger Wörter, semantische Relationen zwischen Deskriptoren, Angaben zu Teilelementen von Komposita und Festen Wendungen - lassen sich in GOLEM in Form von Deskriptorzusätzen (Angabe von Bedeutungsdifferenzierungsnummern), von zusammengesetzten Deskriptoren und durch die Angabe von Beziehungen darstellen (auch hier bietet GOLEM weitergehende Möglichkeiten als TELDOK).

Die syntaktischen Informationen, in denen sich die Satzstruktur widerspiegelt, können in GOLEM zum einen durch Bildung von Komplexen Deskriptoren (d.h. Wortformen, die durch die im Text gegebene syntaktische Relation aufeinander bezogen und verknüpft werden), zum anderen durch Hinzufügung eines Index an den Deskriptor (dies ist die weiterreichende Möglichkeit, da darin mehr syntaktische Informationen dargestellt werden können) auf die DB-Struktur abgebildet werden. (Eine dem Index in GOLEM vergleichbare Möglichkeit, syntaktische Strukturen abzubilden, ist in TELDOK ebenfalls nicht gegeben).

Durch die Realisierung einer GOLEM-Retrievalkomponente sollen also - zusammenfassend dargestellt - folgende Möglichkeiten aufgezeigt und geprüft werden:

Durch Einbeziehung semantischer Informationen in die Recherche soll einerseits die Precision (z.B. durch Verwendung semantisch eindeutiger Deskriptoren) und andererseits der Recall, d.h. die Ermittlung relevanter Dokumente (durch Einbeziehung semantisch und morphosyntaktisch relationierter Begriffe in die Recherche) erhöht werden. Die Nutzbarmachung syntaktischer Strukturen im Dokument für die inhaltliche Erschließung durch Deskriptoren und deren Ver-

wendung bei der Recherche soll ebenfalls zur Verbesserung der Precision beitragen.

#### II.6.1.3 Datenbankaufbau

Eine GOLEM-Datenbank kann aus beliebig vielen so genannten 'Pools', d.h. während der Recherche befragbaren Informationsbanken, bestehen. Ein Pool selbst wiederum besteht aus vier einzelnen Dateien, deren Einrichtung und Strukturierung durch das Programmsystem GOLEM übernommen wird. Auf diesen Dateien operiert das Programmsystem beim Laden der Datenbank und bei der Recherche. Während des Recherchedialogs kann der Pool (auch unter Beibehaltung der Suchparameter) gewechselt werden.

GOLEM-Pooldateien (vgl. SIEMENS 1981) sind im einzelnen:

- O-Datei (Organisationsdatei)
- W-Datei (Wörterbuch/Thesaurus)
- I-Datei (Zielinformationsadressen)
- Z-Datei (Zielinformationsdatei)

Die Abspeicherung der Dokumente erfolgt in Form von Zielinformationen. Eine Zielinformation besteht aus einem Deskriptorenteil und einem Textteil. Text und zugehörende Deskriptoren bilden somit eine Einheit und können nur zusammen abgespeichert werden.

Der Textteil einer Zielinformation enthält den Text des recherchierbaren Dokuments (und ist damit das eigentliche Dokument), während im Deskriptorenteil die zugehörigen Deskriptoren enthalten sind. (Wenn im weiteren von Dokumenten die Rede ist, sind, falls explizit nichts anderes vermerkt ist, Zielinformationen gemeint.)

Außer den Zielinformationen kann (muss aber nicht) auch ein 'Wörterbuch' in die Datenbank eingetragen werden. Zusammen mit den Deskriptoren wird daraus der 'Thesaurus' der Datenbank erstellt. Man beachte hier die unterschiedliche Terminologie: der im System CTX gebräuchliche Ausdruck 'Thesaurus' für ein Verzeichnis semantisch relationierter Begriffe ist in der GOLEM-Terminologie ein Teil des 'Wörterbuchs' (Teil deswegen, da die Relationen oder besser Beziehungen der Begriffe im GOLEM-Wörterbuch ganz allgemeiner Art sein können), ein 'Thesaurus' im GOLEM-Sinne konstituiert sich aus der Zusammenfassung der Deskriptoren der Zielinformationen mit denen des Wörterbuchs.

Wird kein GOLEM-Wörterbuch eingetragen, so besteht der GOLEM-Thesaurus nur aus der Gesamtheit aller im Deskriptorenteil der Zielinformationen vorkommenden Deskriptoren.

##### II.6.1.3.1 Das GOLEM-Wörterbuch

In dem GOLEM-Wörterbuch werden die relationierten Begriffe (Relationendaten) der CTX-Anwendung (hier: Datenschutzrecht) gespeichert. Dabei werden die im CTX-Thesaurus enthaltenen Begriffe und Relationen nicht 1:1 in das Wörterbuch abgebildet, obwohl dies prinzipiell möglich wäre. GOLEM kann - wie erwähnt - bis zu 127 Relationen ("Beziehungen" in der GOLEM-Terminologie) verarbeiten, wobei nur die Synonymierelation vorgegeben ist (die Synonymierelation wird zugleich automatisch invertiert, d.h. es wird eine zweiseitige Beziehung der Begriffe unter-

einander hergestellt; systemseitig werden jedoch automatisch keine Beziehungsketten aufgebaut, also die Transitivität der Synonymie nicht ausgenutzt). Alle anderen Beziehungen sind frei definierbar und werden als einseitige Beziehungen aufgenommen (eine eventuell sinnvolle Invertierung der Relationen muss also vor dem Eintrag in das GOLEM-Wörterbuch vorgenommen werden).

Die Beziehungen werden bei der Einrichtung des 'Pools', d.h. einer GOLEM-Informationsbank, mindestens jedoch vor einem entsprechenden Wörterbucheintrag (wenn es sich dabei um Relationen handelt) vereinbart, wobei die Bezeichnungen für die Beziehungen mit höchstens zwei Zeichen abgekürzt werden können und die zugeordneten Langformen nicht länger als 14 Zeichen sein dürfen. Es sei darauf hingewiesen, dass diese Relationen natürlich keine "semantischen" Relationen (im linguistischen Sinn) zu sein brauchen (vgl. die Relation ID, die auf eine Mehrdeutigkeit hinweist, s.u.).

In der Testanwendung "Datenschutzrecht" sind bisher folgende Beziehungen vereinbart (links stehen die GOLEM-Abkürzungen, in Klammern rechts die im CTX-Thesaurus gebräuchlichen Abkürzungen):

| | | |
|---|---|---|
| S | Synonymie | (SYN) |
| OB | Oberbegriff | (OBR) |
| UB | Unterbegriff | (UNT) |
| AS | Assoziation | (ASS) |
| QU | Quasisynonymie | (QUA) |
| TE | Teil | (TEI) |
| GA | Ganzes | (GAN) |
| RE | juristischer Regelungsgegenstand | (REG) |
| IU | juristische Regelung | (IUS) |
| NE | Nebenbegriff | (NEB) |
| GE | Gegenbegriff | (GEG) |
| AB | Abkürzung | (ABK) |
| LN | Langform | (LNF) |
| ID | Informationsdokument | - |

(Wegen der Beschränkung auf 14 Zeichen bei den Langformen wurden bei der Beziehungsvereinbarung in GOLEM die folgenden Abkürzungen gewählt: RE=j.Regelggstd, IU=jur.Regelung, ID=Infodokument.)

Unter Synonymie sind eine ganze Reihe von speziellen Synonymierelationen des CTX-Thesaurus (Kap. II.3.6) zusammengefasst; z.B. die Synonymierelationen zwischen Rechtschreibvarianten (SYS), zwischen Nominalkomposita und den daraus abgeleiteten mehrwortigen Begriffen (SKM) bzw. zwischen mehrwortigen Begriffen (und Festen Wendungen) und den daraus abgeleiteten Nominalkomposita (SMK, SFK), zwischen mehrwortigen Begriffen und den entsprechenden Komplexen Deskriptoren (z.B. SMG, SMA), wobei noch zwischen den syntaktischen Relationen, die in dem mehrwortigen Begriff enthalten sind, differenziert wurde. Die Zusammenfassung der im CTX-System aus verschiedenen Gründen - auch zu Testzwecken - weiter differenzierten speziellen Synonymierelationen erfolgte aus der Überlegung heraus, dass zum Einen die Handhabung vieler sich nur wenig unterscheidender Relationen beim Retrieval unübersichtlich wird und dass zum Anderen die feine Differenzierung der Synonymierelationen zwar für die Erstellung

und Erweiterung des CTX-Thesaurus benötigt wird, für ein Retrieval jedoch kaum inhaltliche Verbesserung bringen dürfte. Weiterhin wurden die im CTX-Thesaurus differenzierten Derivationsrelationen in der GOLEM-Variante auf die S-Relation abgebildet.

Begriffe, die durch die Relationen AB oder LN verknüpft sind, sind zusätzlich - also neben der spezifischen Relationierung - durch die Synonymierelation verknüpft worden, da die Relation zwischen einer Abkürzung und der entsprechenden Langform als eine Synonymierelation (sozusagen als eine besondere Schreibvariante) aufgefasst werden kann. Der Benutzer ist somit in der Lage, auch nach dem Ausschalten der Synonymierelation im Retrieval noch die Abkürzungen und die Langformen in eine Recherche miteinzubeziehen, und umgekehrt muss er bei eingeschalteter Synonymierelation nicht noch die Relationen AB und LN zuschalten.

Die Relation ID stellt keine semantische Relation dar. Sie ist eingeführt, um die Verbindung eines mehrdeutigen Lemmas mit einem besonderen Dokument (genannt Informations- oder Infodokument) herzustellen, in dem die Vereindeutigung durch Bedeutungsnummern an Hand von Beispielen und Paraphrasen erklärt ist.

Diese Informationsdokumente werden bei CTX-II über eine spezielle Software automatisch aus den Angaben des Differenzierungswörterbuchs erzeugt.

#### II.6.1.3.2 Zielinformationen

Es liegen zwei Arten von Zielinformationen vor:

> (1) Zielinformationen, deren Textteil den eigentlichen Recherchegegenstand (Normen, Entscheidungen, etc.) enthalten;
>
> (2) Zielinformationen, deren Textteil Informationen zu mehrdeutigen Deskriptoren enthält, bzw. auch Definitionen oder Quellenangaben (Infodokument). Diese zweite Art von Zielinformationen dient bei der Recherche als Hilfsmittel. Damit kann sich der Benutzer z.B. über die für seine Zwecke richtige Bedeutungsdifferenzierungsnummer bei mehrdeutigen Deskriptoren informieren.

zu (1): Zielinformationen zu inhaltstragenden Dokumenten

Diese Zielinformationen bestehen aus zwei Deskriptorenabschnitten und einem Textabschnitt.

Der erste Deskriptorenabschnitt enthält einen zusammengesetzten Deskriptor mit dem Aspekt "E-DAT"; der gebundene Deskriptor zu diesem Aspekt besteht aus der Ziffernfolge des Einspeicherungsdatums (Beispiel: E-DAT:811102 bedeutet, dass die Zielinformation am 2.11.1981 eingespeichert wurde). Dieser zusammengesetzte Deskriptor wird zu jedem Dokument beim Eintragen der Zielinformation in die Datenbank erzeugt.

Der zweite Deskriptorenabschnitt enthält die das Dokument inhaltlich erschließenden Deskriptoren. Dies können freie und zusammengesetzte Deskriptoren sein. Sie werden in ihrer im Text vorgegebenen Reihenfolge aufgeführt, d.h. nicht alphabetisch sortiert. Diese Darstellungsart ist für Test- und Auswertungszwecke gewählt worden. Für die Deskriptoren gilt eine Längenbeschränkung von 255 Zeichen eines definierten Zeichenvorrats. Es dürfen in der Deskriptorkette

nicht beliebige Zeichenkombinationen auftreten, da manche als Steuerzeichen reserviert sind, wie z.B. die Zeichen "((" und "(*"; eine Beschränkung, die sich praktisch allerdings nicht auswirkt.

Grundlage zur Erstellung der Deskriptoren des zweiten Deskriptorenabschnitts sind die Ergebnisse der automatischen Textanalyse, die in Form von Deskriptorbasisdaten mithilfe des Programmes zur Deskriptorermittlung in eine weitestgehend maschinenunabhängige Schnittstelle umgesetzt sind und als Eingabedaten zur automatischen Erstellung der Zielinformationen zur Verfügung stehen. Die Umsetzung in das erforderliche GOLEM-Eingabeformat erfolgt mithilfe des Programms DUMSG (=Daten-UMSetzprogramm GOLEM).

Die Deskriptoren werden dabei mit weiteren Informationen in Form von Indices versehen; dies dient der Darstellung der syntaktischen Informationen, die Auskünfte über die "syntaktische Rolle" des Deskriptors im Dokument geben.

zu (2): Informationsdokumente

Diese Zielinformationen bestehen ebenfalls aus zwei Deskriptorenabschnitten und einem Textabschnitt.

Im ersten Deskriptorenabschnitt steht bei allen diesen Zielinformationen der Deskriptor INFORMATIONSDOKUMENT. Er wird als Wahldeskriptor bei der Einspeicherung generiert (d.h. zusätzlich zum Deskriptorenbestand eines Dokuments aufgenommen) und ermöglicht ein einfaches Identifizieren aller Informationsdokumente. Der zweite Deskriptor im ersten Deskriptorenabschnitt besteht aus einem zusammengesetzten Deskriptor mit dem Aspekt E-DAT, der das Einspeicherungsdatum angibt.

Der zweite Deskriptorenabschnitt besteht aus einem zusammengesetzten Deskriptor mit dem Aspekt INFO. Der gebundene Deskriptor zu diesem Aspekt besteht aus dem in diesem Dokument erklärten Formaldeskriptor (d.h. einem mehrdeutigen Deskriptor, Beispiel: INFO:ANLAGE ist der zusammengesetzte Deskriptor, der zu dem Dokument führt, in dem das mehrdeutige Wort ANLAGE disambiguiert und durch Paraphrasen erklärt ist). Da auf diese Weise alle mehrdeutigen Wörter aspektiert sind, kann man sich in der GOLEM-Version schnell einen Überblick über alle oder auch nur einige mehrdeutige Lemmata verschaffen. Das geschieht durch die Ausgabe aller oder eines Bereichs der gebundenen Deskriptoren zum Aspekt INFO aus dem Aspekt-Thesaurus (Beispiel: mit dem Kommando E INFO erhält der Benutzer eine Auflistung aller an den Aspekt INFO gebundenen Deskriptoren und damit eine Übersicht zu allen (vorhandenen) mehrdeutigen Begriffen).

Über die Beziehungsart (bzw. Relation) ID ist eine Verknüpfung zwischen diesen zusammengesetzten Deskriptoren und den aus den Wortlauten der mehrdeutigen Wörter bestehenden freien Deskriptoren (Formaldeskriptoren) hergestellt (Beispiel: ANLAGE ist im Thesaurus durch die Beziehung ID mit dem zusammengesetzten Deskriptor INFO:ANLAGE verknüpft).

Bei einer Recherche bietet sich dann folgendes Vorgehen an: Da der Benutzer beim Aufbau der Suchwortliste in der Regel nicht wissen wird, welche Deskriptoren mehrdeutig sind (er kann es i.a. nur vermuten), das System (im Gegensatz zu TELDOK) aber auch nicht über die Möglichkeit verfügt, bei einer Suchanfrage mit einer bestimmten Sorte von Deskriptoren (bei TELDOK den mit der Relation Homonymie markierten) automatisch einen Hinweis an den Benutzer zu geben,

dass er dabei ist, mit einem mehrdeutigen Deskriptor zu recherchieren, wird nach dem Aufbau der Suchwortliste ein weiteres Kommando nötig (s. Beispiel unten), mit dem die Suchwortliste um die Begriffe erweitert wird, die mit der Beziehung ID zu den Deskriptoren in der Suchwortliste relationiert sind. Aus der so erweiterten Suchwortliste entnimmt dann der Benutzer, welche der Deskriptoren mehrdeutig sind, und gleichzeitig, mit welchen Deskriptoren er sich Zugang zu den betreffenden Informationsdokumenten verschaffen kann.

Dazu ein kleiner Beispieldialog (die Suchwortliste besteht nur aus einem Deskriptor):

1 ANLAGE
(die Zahl vor dem Deskriptor identifiziert ihn in der Suchwortliste und muss bei der weiteren Recherche, die dann erst zu den Dokumenten führt, verwendet werden).

Der Benutzer erweitert nun die Suchwortliste um die mit der Beziehung ID relationierten Begriffe, indem er das entsprechende Kommando eingibt:

GBEZ */ID
Die erweiterte Suchwortliste erscheint wie folgt:

1 ANLAGE
2 ID...INFODOKUMENT....
3 INFO:ANLAGE

Will der Benutzer das Informationsdokument sehen, so braucht er nur noch mit der Deskriptorennummer 3 (Kommando L 3) zu recherchieren. Als Alternative bietet sich an, die Relation über ein bei GOLEM verfügbares Voreinstellungskommando systematisch in die Recherche einzubinden. Mit dem Kommando SGEN ID kann somit der gleiche Effekt unmittelbar bei der Recherche mit einem mehrdeutigen Wort erreicht werden.

Die oben beschriebenen Zielinformationen werden strukturiert in die Datenbank eingegeben. Grundlage zur automatischen Erstellung dieser Zielinformationen sind - wie erwähnt - die im Differenzierungswörterbuch gesammelten Informationen über mehrdeutige Wörter.

Harald H. Zimmermann, Edith Kroupa, Gerald Keil (1983):
C T X - Ein Verfahren zur computergestützten Texterschließung
TEIL III

### II.6.1.3.3 Deskriptoren in der GOLEM-Anwendung

- Zusammengesetzte Deskriptoren

Zusammengesetzte Deskriptoren bestehen (in der GOLEM-Terminologie) aus einem Aspekt und einem gebundenen Deskriptor. Es wurden (zusätzlich zum vorgegebenen GOLEM-Aspekt E-DAT) in der GOLEM-Datenbank u.a. folgende Aspekte eingeführt (vgl. Kap. II.5):

| | |
|---|---|
| INFO | weist auf Informationsdokumente hin |
| ART | Textart |
| DATUM | Veröffentlichungsdatum |
| GLIEDERUNG | Titel eines Textkapitels |
| FUNDSTELLE | formales Ablagekriterium |
| HAUPTTITEL | Titel des gesamten Textes |
| HERAUSGEBER | Herausgeber des Textes |
| TITEL | Schlagwörter aus dem Titel |
| TYP | Texttyp (z.B. Tätigkeitsbericht, Norm) |
| VERFASSER | Verfasser des Textes |

- Freie Deskriptoren

In der CTX-Terminologie wird unterschieden zwischen Einfachen Deskriptoren und Komplexen Deskriptoren. Beide Arten werden in der GOLEM-Realisierung als freie Deskriptoren abgespeichert.

Die Einfachen Deskriptoren bestehen nur aus dem Wortlaut, der in den Fällen, in denen dieser mehrdeutig ist und eine Vereindeutigung durchgeführt werden konnte, um eine so genannte Bedeutungsdifferenzierungsnummer ergänzt wird (Beispiel: DATUM1).

Wortformen, die den Analysewortklassen Substantiv (SUB), Verb (VRB), Adjektiv (ADJ), angehören (wobei die aktuelle, d.h. im Text ermittelte Analysewortklasse zugrunde gelegt wird), werden in ihrer lexikalischen Form (Grundform) als freie Deskriptoren vergeben. Außerdem sind noch Eigennamen (sie werden z.Zt. auf die Wortklasse SUB=Substantiv abgebildet) und Ziffernfolgen (Wortklasse NUM =Zahlwort) als Deskriptoren zugelassen.

Phrasen, die als Feste Wendung erkannt werden bzw. ADJ-Relationen, werden ebenfalls als freie Deskriptoren vergeben (Beispiel: PERSONENBEZOGENE DATEN, AUSSER KRAFT SETZEN).

Stehen zwei Deskriptoren in einer der syntaktischen Relationen Genitivattribut (G), Präpositionalattribut (P) oder Anreihung (K), so wurde durch CTX ein so genannter Komplexer Deskriptor erzeugt. Er besteht auch in der GOLEM-Version aus den einzelnen Wörtern, getrennt durch je ein Leerzeichen, und der Abkürzung der entsprechenden syntaktischen Relation (Beispiel: BEHOERDE K STELLE; SCHUTZ G DATUM).

Auch wenn die in einer syntaktischen Relation stehenden (einfachen) Wörter mehrdeutig sind, entfällt die Bedeutungsdifferenzierungsnummer bei der Erstellung des Komplexen Deskriptors, da davon ausgegangen wird, dass in dem durch den Komplexen Deskriptor hergestellten Kontext eine ausreichende Eindeutigkeit gewährleistet ist (Beispiel: ANLAGE P GESETZ).

- Indizierung und Feinrecherche

GOLEM bietet - wie erwähnt - die Möglichkeit, zu Deskriptoren so genannte "Indices" anzugeben, deren Informationen in der "Feinrecherche" ausgenutzt werden. Ein Index besteht aus Rollenindikatoren und Indexziffern. Rollenindikatoren und Indexziffern können dergestalt aufeinander bezogen werden, dass die Indexziffern, die in der Reihenfolge vor einem Rollenindikator

angeordnet sind, sich als "Thema" auf diesen Rollenindikator beziehen. Die Länge eines Index darf 255 Zeichen nicht überschreiten, die Indexziffern müssen kleiner oder gleich 255 sein. Der Index wird im Rahmen des CTX-Systems dazu genutzt, einen Teil der syntaktischen Informationen, die die automatische Textanalyse liefert, auf eine Folge von Indexziffern und Rollenindikatoren abzubilden. Die zur Verfügung stehenden syntaktischen Informationen beinhalten im wesentlichen Angaben zur Zugehörigkeit eines oder mehrerer Deskriptoren zu einer (d.h. derselben) syntaktischen Einheit wie Satz, Subsatz, Nominalgruppe und Verbalgruppe, sowie Wortklassenangaben und Angaben über syntaktische Relationen.

Prinzipiell eignen sich alle durch die syntaktische Analyse ermittelten Einheiten und Relationen als Rollenindikatoren, jedoch wird in der Testphase gegenwärtig nur ein Teil davon verwendet. Durch die zusätzliche Angabe von Indexziffern zu den Rollenindikatoren werden die Deskriptoren wie folgt aufeinander bezogen: Den Deskriptoren werden im Index die Rollenindikatoren zugeordnet, die auf sie zutreffen. So kommt z.B. ein Deskriptor jeweils in einem Satz und einem Subsatz vor, er gehört einer Wortklasse an und steht darüber hinaus eventuell in einer der oben angeführten syntaktischen Relationen zu anderen Deskriptoren. Gehören nun zwei (oder auch mehrere) Deskriptoren derselben syntaktischen Einheit (z.B. Satz oder Subsatz) an oder stehen sie in derselben syntaktischen Relation, so erhalten sie vor dem entsprechenden Rollenindikator dieselbe Indexziffer.

Die Indexzahlen sind als Themenbereiche aufzufassen, denen die Deskriptoren in Abhängigkeit von ihren Rollenindikatoren (dokumentweise) zugeordnet werden.

Als Beispiel dazu soll der folgende Satz dienen:

"Aufgabe des Datenschutzes ist der Schutz personenbezogener Daten vor Missbrauch."

Als Deskriptoren dazu sollen ermittelt sein (es werden zur Vereinfachung Feste Wendungen in dem Beispiel nicht berücksichtigt, ebensowenig Bedeutungsdifferenzierungen und Teildeskriptoren):

AUFGABE
DATENSCHUTZ
SCHUTZ
PERSONENBEZOGEN
DATUM
MISSBRAUCH

Als Rollenindikatoren seien gegeben:

(1) die syntaktischen Einheiten

| | |
|---|---|
| SATZ | (S) |
| SUBSATZ | (SB) |
| WORTKLASSE | (SUB, ADJ) |

(2) die syntaktischen Relationen

'linkes' bzw. 'rechtes Element' einer Genitivrelation (GL, GR)
'linkes' bzw. 'rechtes Element' einer PRP-Relation (PL,PR)

Alle Deskriptoren kommen in demselben Satz (und zugleich Subsatz) vor; den entsprechenden Rollenindikatoren werden also jeweils dieselben Indexziffern (hier 1 für Satz = S und 51 für Subsatz = SB) vorangestellt. AUFGABE und DATENSCHUTZ kommen darüber hinaus in einer Genitiv-Relation vor, und zwar hat in dieser Relation AUFGABE die Rolle des linken und DATENSCHUTZ die Rolle des rechten Elements. AUFGABE erhält deshalb den Rollenindikator GL und DATENSCHUTZ den Rollenindikator GR.

Da beide Deskriptoren in derselben Relation vorkommen, erhalten beide Deskriptoren vor dem entsprechenden Rollenindikator dieselbe Indexziffer (hier: 121).

Ebenso verhält es sich bei den Deskriptoren SCHUTZ und DATUM. Da diese beiden Deskriptoren jedoch in einer anderen Genitivrelation vorkommen als AUFGABE und DATENSCHUTZ, erhalten sie zwar gleiche Indexziffern, aber eben andere als die, die vor den entsprechenden Rollenindikatoren von AUFGABE und DATENSCHUTZ stehen (hier: 122).

Deskriptoren, die in einer Adjektiv-Substantiv-Relation stehen, erhalten zusätzlich eine dem Rollenindikator "Wortklasse" zugeordnete Indexziffer, und zwar dieselbe, die vor dem Rollenindikator 'A' steht (hier: 123). Damit können Elemente dieser Relation über die Wortklasse und über die Angabe der syntaktischen Relation zusammengeführt werden. Die Unterscheidung bei Genitiv- und Präpositionalrelationen in Links- und Rechtselement ist notwendig, da die Deskriptoren nicht als geordnetes Paar auftreten und diese Relationen nicht symmetrisch sind (bei der A-Relation kann eine Verwechslung wegen der verschiedenen Wortklassen nicht vorkommen).

Die Deskriptoren und ihre Indices haben also folgende Gestalt (dabei spielt der numerische Wert an sich keine Rolle, lediglich der gleiche Wert vor entsprechenden Rollenindikatoren identifiziert die jeweilige Zusammengehörigkeit):

AUFGABE (1,S,51,SB,SUB,121,GL)
DATENSCHUTZ (1,S,51,SB,SUB,121,GR)
SCHUTZ (1,S,51,SB,SUB,122,GL)
DATUM (1,S,51,SB,SUB,122,GR)
PERSONENBEZOGEN (1,S,51,SB,123,ADJ,123,A)
DATUM (1,S,51,SB,123,SUB,123,A)
SCHUTZ (1,5,51, SB,SUB,124,PL)
MISSBRAUCH (1,S,51,SB,SUB,124,PR).

Dem Benutzer stehen also zwei Suchstrategien offen:

- einerseits kann er sich über die 'Grobrecherche' durch die Angabe Komplexer Deskriptoren Zugang zu den gesuchten Dokumenten verschaffen (Komplexe Deskriptoren als Bild der im Text aktuellen syntaktischen Bezogenheit der Begriffe aufeinander),
- andererseits kann er über die (in GOLEM etwas umständliche) "Feinrecherche" eine für die Präzisierung der Recherche relevant erscheinende syntaktische Beziehung der Einfachen Deskriptoren untereinander abfragen und so die in der vorausgehenden "Grobrecherche" angelieferten Dokumente weiter einschränken.

- Da auch den Komplexen Deskriptoren noch ein Index angefügt wird, der die Rollenindikatoren SATZ und SUBSATZ beinhaltet, eröffnet sich v.a. die Möglichkeit, sich auch neben einer Recherche mit Komplexen Deskriptoren noch des Mittels der Feinrecherche zu bedienen, um die Ergebnisse der Recherche im Hinblick auf das Vorkommen in der gleichen syntaktischen Einheit zu präzisieren, etwa für den Fall, dass nur solche Dokumente als Ergebnis gewünscht werden, in denen angegebene Komplexe Deskriptoren im selben Satz oder im selben Subsatz vorkommen.

An dieser Stelle ist zu erwähnen, dass durch den Aufbau der Komplexen Deskriptoren und die Darstellung syntaktischer Relationen im Index von Einfachen Deskriptoren die eingangs erwähnten unterschiedlichen Möglichkeiten der Abbildung der Indexierungsergebnisse auf die DB-Struktur realisiert wird.

- Textteil

Der Textteil der (Text-)Dokumente besteht aus einem Textabschnitt. Bei der Erstellung der Zielinformationen wird derzeit an den Anfang eines jeden Textabschnitts eine Kennung gesetzt, die angibt, um welches Dokument es sich handelt. Die Kennung besteht aus dem Wortlaut des an den Aspekt "NR" gebundenen Deskriptors der jeweiligen Zielinformation, also z.B. BDSG, und es wird eine interne Dokumentnummer (z.B. 007) erstellt.

Beispiel: NR: BDSG 007 besagt, dass es sich um das 7. Dokument (nicht notwendig um den 7. Paragraphen) des Bundesdatenschutzgesetzes handelt.

Diese Kennung und die interne Dokumentnummer sind daneben als zusammengesetzte bzw. als freie Deskriptoren vergeben, so dass es möglich ist, ohne Weiteres die einem gefundenen Dokument vorhergehenden oder die darauf folgenden Dokumente aufzufinden.

Da, wie schon erwähnt, bestimmte Zeichen und Zeichenkombinationen als Steuerzeichen reserviert sind, ist der Textteil vor der Eingabe in die Informationsbank daraufhin zu überprüfen, oh er solche Zeichen enthält. Dies kann im Zusammenhang mit einer Präedierung des Textes geschehen.

### II.6.1.4 Erste Tests

Bisher wurde nur zu Test- und Präsentationszwecken in der GOLEM-Umsetzung recherchiert. Dabei wurden die Recherchen in der Regel von Projektmitarbeitern durchgeführt, die sowohl mit dem Inhalt der Informationsbank, als auch mit entsprechenden Formulierungs- und Abfragemöglichkeiten vertraut waren. Insofern konnte in der Laboranwendung kein anwenderbezogener Test durchgeführt werden. Dennoch soll hier auf erste projektinterne Erfahrungen hingewiesen werden.

Die Einführung syntaktischer Kategorien als Mittel zur Feinrecherche bewährt sich. Der Rollenindikator SATZ ist dabei eher als eine Ergänzung zu den anderen derzeit verwendeten Rollenindikatoren wie SUBSATZ oder WORTKLASSE zu sehen; in Texten, in denen wie z.B. im BDSG sehr lange Sätze vorkommen, ist er jedoch kaum noch ein Präzisierungswerkzeug. Besser geeignet sind feinere Einheiten wie SUBSATZ oder NOMINALGRUPPE. Die Wortklasse kann dazu beitragen, Homographen zu vereindeutigen. Ein Beispiel dazu: KOSTEN kann durch Angabe des

Rollenindikators SUBSTANTIV auf eine Bedeutung reduziert werden (damit werden die Dokumente, in denen KOSTEN als Verb auftritt, ausgeschieden).

Da die Nutzung der syntaktischen Relationen für die Recherche auf zwei Ebenen (Grobrecherche mit Komplexen Deskriptoren, Feinrecherche mit Indices) möglich ist, konnte hier ein direkter Vergleich durchgeführt werden. Dabei zeigten sich deutlich die besonderen Möglichkeiten, aber auch einige derzeit noch gegebene Schwächen des Systems: Ein Beispiel sind die Komplexen Deskriptoren und Festen Wendungen. Bei der Grobrecherche mit derartigen Begriffen ist eine hohe Precision erreichbar, doch ist die Qualität von der erreichten Analysetiefe und der Korrektheit der automatischen Satzanalyse abhängig. Dies wirkt sich vor allem in einem verminderten Recall aus, wenn auf der Grobrechercheebene "nur" mit Komplexen Deskriptoren recherchiert wird. Dokumente, deren Analyse nicht "tief" genug gelungen ist, um die Nominalanalyse durchzuführen, d.h. deren syntaktische Relationen nicht erkannt worden sind, werden bei einer solchen Recherche nicht nachgewiesen, obgleich die Einzeldeskriptoren richtig ermittelt wurden, ebenso andere syntaktische Einheiten (z.B. Subsatzgrenzen). Versucht man die syntaktischen Relationen bei der Feinrecherche zur Präzisierung auszunutzen, d.h. wird verglichen, ob die Indices der Einfachen Deskriptoren unter einem bestimmten Gesichtspunkt (hier: syntaktische Rolle) übereinstimmen, wird zwar ein Dokument, dessen Nominalstruktur nicht analysiert worden ist, auch nicht gefunden, jedoch ist der Benutzer darüber informiert, dass es mehr - als durch die Feinrecherche ausgewiesenen - Dokumente gibt, in denen alle angegebenen Einfachen Deskriptoren vorkommen. Er könnte dann daraufhin entscheiden, ob er eine Feinrecherche mit gröberem Raster (etwa "Vorkommen im gleichen Satz") anschließen will. Dieses Problem lässt sich jedoch auch dadurch beheben, dass z.B. durch geeignete Postedition der Texte in allen Fällen sichergestellt wird, dass jeder Satz zumindest die Nominalgruppenanalyse erfolgreich durchläuft.

Ein anderes Problem in diesem Zusammenhang, das Auswirkungen auf die zu wählende Recherchestrategie - Grob- oder Feinrecherche - hat, ergibt sich daraus, dass bei der Verwendung syntaktischer Relationen in der Recherche nur die Dokumente gefunden werden, in deren Texten die angegebenen syntaktischen Relationen auch belegt sind. Unter Umständen werden also Dokumente nicht nachgewiesen, die der Benutzer unter dem Gesichtspunkt seines Problems durchaus als relevant bezeichnen würde. Ein Beispiel möge dies veranschaulichen. Es sollen Dokumente gefunden werden zum Problem "Schutz der Daten bei ihrer Verarbeitung". Beginnt man die Recherche mit dem Komplexen Deskriptor SCHUTZ G DATUM und dem einfachen Deskriptor VERARBEITUNG, so wird ein Dokument, in dem die Phrase "Schutz vor Missbrauch der Daten bei ihrer Verarbeitung" vorkommt, nicht angezeigt, da darin DATUM nicht Genitivattribut zu SCHUTZ, sondern Genitivattribut zu MISSBRAUCH ist (es liegt also kein Systemfehler vor). Jedoch könnte ein Benutzer auch ein Dokument mit obigem Inhalt als (für sein Problem) relevant betrachten. Die zweite Möglichkeit, nämlich die Ausnutzung der syntaktischen Relationen in der Feinrecherche, setzt eine Recherche in obigem Beispiel mit den Einfachen Deskriptoren SCHUTZ, DATUM und VERARBEITUNG voraus. Dabei wird der Benutzer über die Anzahl der Dokumente informiert, in denen alle drei Suchbegriffe vorkommen. Daran schließt sich die Feinrecherche an, in der die Deskriptoren SCHUTZ und DATUM daraufhin verglichen werden, ob sie in der syntaktischen Relation G zueinander stehen. Dabei wird dann zwar das obige Dokument auch nicht nachgewiesen, jedoch weiß der Benutzer für den Fall, dass er mithilfe der ihm daraufhin gezeigten Dokumente sein Problem nicht vollständig lösen kann, dass er noch mehr Dokumente, in denen alle drei Suchbegriffe vorkommen, zur Verfügung hat. Er kann dann - wie oben bereits vorgeschlagen - eventuell eine weitere Feinrecherche auf Satz- oder Subsatzebene anschließen.

Der Nachteil der Feinrecherche liegt jedoch in ihrer Unhandlichkeit und dem gesteigerten Rechenaufwand, der sich bei einer größeren Dokumentmenge in einer längeren Wartezeit bemerkbar macht.

Beim Zuschalten semantischer Begriffsbeziehungen bei der Recherche mit Einfachen Deskriptoren hat es sich gezeigt, dass wegen der Vielzahl der in der Synonymierelation enthaltenen speziellen semantischen Relationen z.T. auch unerwünschte Dokumente gefunden wurden, deren Auffinden auf einen in einer Synonymierelation zum Suchbegriff stehenden Begriff zurückzuführen ist, den der Benutzer nicht darin erwartet hätte. Durch eine verbesserte Thesaurusstrukturierung kann dieses Problem jedoch in der praktischen Anwendung schrittweise bereinigt werden.

### II.6.2 Umsetzung TELDOK

#### II.6.2.1 Allgemeines

In der ersten Projektphase wurde (in Regensburg) eine TELDOK-Implementierung (bezogen auf eine Teilmenge der jetzt verfügbaren Daten) vorgenommen. Hierbei zeigte sich besonders die Problematik der Anpassung an bestehende Retrievalsysteme. Obgleich die Ergebnisse dieser Implementierung JUDO-T inzwischen schon drei Jahre zurückreichen, sind die Erfahrungen m.E. so wichtig, dass sie hier zusammengefasst vorgestellt werden sollen.

TELDOK ist das Information-Retrieval-System, das vom Hersteller des TR 440-Rechners angeboten wird (vgl. TELDOK 1978). Obwohl diesem System (etwa im Unterschied zu GOLEM) eine flexible Datenmanipulationssprache (das Datenbanksystem DBS des TR 440) zugrunde liegt, werden vom TELDOK-Programmkomplex nur bestimmte Datenstrukturen zugelassen und beim Retrieval unterstützt. Diese Einschränkungen wirkten sich - wie sich zeigen wird - zum Teil gravierend aus. Es sind hier vor allem zu nennen

- die Längenbeschränkung für Sätze, (d.i. hier: Deskriptoren) die im Direktzugriff gehalten werden sollen, auf maximal 29 Zeichen,
- die Anzahl der über Deskriptoren definierbaren Relationen: sie ist auf 6 beschränkt; zudem sind diese Relationen bereits begrifflich vordefiniert.

#### II.6.2.2 TELDOK-Strukturen

Ebenso wie die GOLEM-Anwendung von CTX (vgl. Kap. II.6.1), stellte sich auch bei TELDOK das Problem, möglichst viele der in den anwendungsunabhängigen Schnittstellen vorliegenden Informationen in TELDOK abzubilden. Das Ergebnis dieser Abbildung sollte dabei für potentielle Benutzer hinsichtlich des Komforts und der Benutzbarkeit akzeptabel bleiben. Durch Umdefinition bestehender TELDOKStrukturen (im Falle der Normal/Formaldeskriptoren) bzw. durch Überlagerung der bestehenden Strukturen mit zusätzlichen Datenstrukturen (unterschiedliche Deskriptoren- und Dokumententypen) wurde diese Abbildung geleistet.

(1) Thesaurusteil

In einer TELDOK-Informationsbank findet eine logische Trennung zwischen Thesaurusteil (inverted file) und Dokumententeil (Pool) statt. Im Thesaurusteil einer solchen Informationsbank

sind realisiert:

- Die Verknüpfung von Deskriptoren mit den Dokumenten, in denen die Deskriptoren belegt sind. Diese Verknüpfung erfolgt über Pointer, die in den Dokumententeil verweisen und vom Datenbanksystem verwaltet werden.
- Die Verknüpfung von Deskriptoren untereinander mittels Pointern, die in den Deskriptorenteil rückverweisen.

Dabei erfolgen diese Verknüpfungen nicht unmittelbar, sondern über eine Zwischenstufe, die die Angabe der Verknüpfungsart ermöglicht. Es sind somit Relationen zwischen Deskriptoren und Dokumenten bzw. zwischen Deskriptoren untereinander benennbar. Die möglichen Relationen über Deskriptoren sind dabei in TELDOK vorgeschrieben; sie werden vom TELDOK-Programmsystem als Thesaurusrelationen behandelt. TELDOK sieht insgesamt 6 solcher Relationen vor.

Die Relationen zwischen Deskriptoren und Dokumenten spezifizieren die Art des Vorkommens eines Deskriptors in einem Dokument. Damit besteht die Möglichkeit, die Deskriptoren in bestimmte inhaltliche Kategorien einzuordnen und diese Kategorien bei einer Suche anzugeben. Solcherart kategorial bestimmte Deskriptoren nennt man in der TELDOK-Technologie gebundene Deskriptoren. Im Gegensatz zu den Relationen zwischen Deskriptoren sind diese Relationen (Kategorien) vom Benutzer frei definierbar. Vom TELDOK-System werden max. 255 solcher Kategorien verwaltet.

Der Thesaurusteil des TELDOK-Systems setzt sich somit aus 5 verschiedenen Satztypen zusammen:

- Deskriptorsätze: dieser Satztyp wird in der Datenbank im Direktzugriff gehalten. Er ermöglicht den indirekten Zugriff auf Dokumente über inhaltliche Kriterien.

- Kategoriensätze: in diesen Sätzen wird die inhaltliche Kategorie des Vorkommens eines Deskriptors angegeben. Besteht in dieser Hinsicht keine Einschränkung, so spricht man von einem freien Deskriptor. In diesem Fall ist der Kategoriensatz leer. Andernfalls heißt der Deskriptor gebunden.

- Dokumentadressensätze: diese Sätze enthalten die Pointer in den Dokumententeil und verbinden den Deskriptor in der jeweiligen Kategorie mit den Dokumenten, in denen er vorkommt.

- Semantische Verknüpfungssätze: dieser Satztyp ermöglicht es, die Relationen, die zwischen Deskriptoren bestehen, anzugeben. Vorgesehen sind dabei von TELDOK folgende Relationen: Synonymie, Antonymie, Oberbegriff, Unterbegriff, semantisches Feld, Homonymie.

- Deskriptoradressensätze: dieser Satztyp enthält die Pointer, die einen Deskriptor mit seinen semantischen Relata verbinden.

Der logische Zusammenhang zwischen diesen Satztypen kommt in dem Kettenstenogramm des

Thesaurusteils zum Ausdruck (vgl. Abb. II.6).

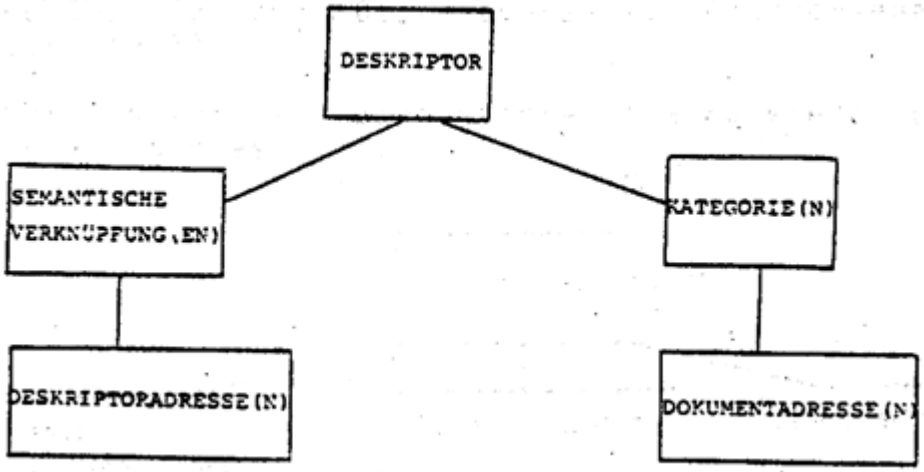



Abb. II.6: Kettenstenogramm des Thesaurusteils von TELDOK.

Diese Datenstruktur unterliegt nun folgenden Restriktionen:

- der Deskriptorsatz darf nur max. 29 Zeichen lang sein,
- die möglichen Werte der semantischen Verknüpfungssätze sind vordefiniert.

(2) Dokumententeil

Der Dokumententeil des TELDOK-Systems enthält zum einen die Informationen, die durch die Deskriptoren des Thesaurusteils erschlossen wurden. Zum anderen findet sich zu jedem Dokument ein Rückverweis auf die in ihm belegten Deskriptoren (in der Terminologie von TELDOK "Hintergrunddeskriptoren"), getrennt nach freien und gebundenen Deskriptoren. Als Zugang zu diesen beiden "Informationspfaden" dient jedesmal der Dokumentschlüssel. Diese ein Dokument innerhalb der Datenbank eindeutig identifizierende Kennung wird, wie die Deskriptorsätze, ebenfalls im Direktzugriff gehalten. Das bedeutet, dass dieser Satztyp ebenso wie die Deskriptoren unmittelbar Argument einer Suche sein kann.

Im Dokumententeil finden sich somit die folgenden Satztypen:

- Dokumentschlüssel: dieser Satztyp kann beliebige alphanumerische Zeichenketten aufnehmen, die jedoch in ihrer Länge wieder auf max. 29 Zeichen beschränkt sind;

- Referatsätze: Referatsätze enthalten satzweise den Dokumententext;

- Freie Deskriptoradressensätze: dieser Satztyp enthält die Pointer, die auf Deskriptoren im Deskriptorenteil verweisen, die ohne Kategorieneinschränkung zu den betreffenden Dokumentenschlüssel belegt sind;

- Gebundene Deskriptoradressensätze: wie oben, jedoch für gebundene Deskriptoren Deskriptoren mit Kategorienangabe im Thesaurusteil).

Die logische Struktur des Dokumententeils wird wieder durch ein Kettenstenogramm dargestellt.

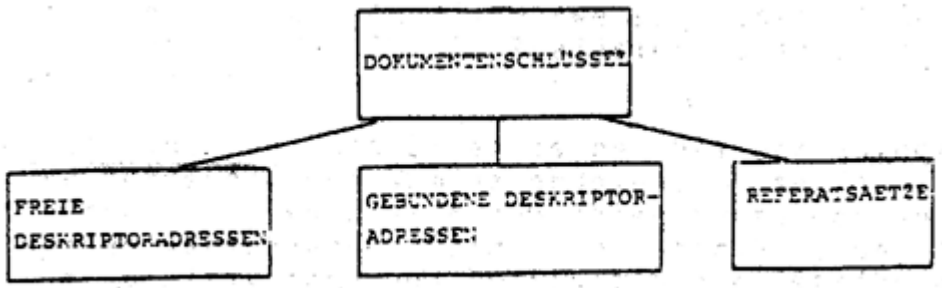



Abb. II.7: Kettenstenogramm des Dokumententeils von TELDOK

### II.6.2.3 Informationen der Schnittstelle

Die in die TELDOK-Variante abzubildenden Daten entstammen z.T. der Schnittstelle des Systems zur Dokumentendeskribierung oder sind das Ergebnis intellektueller lexikalisch-semantischer bzw. fachgebietsbezogen-pragmatischer Analysen. Im einzelnen liegen vor:

- den Inhalt des Textes erschließende Wörter und Wortsyntagmen in lemmatisierter Grundform. Als Ergebnisse der automatischen Deskribierung zerfallen diese Deskriptoren in 4 Klassen:

  (1) Normaldeskriptoren: Damit werden Deskriptoren bezeichnet, die vereindeutigt sind im Sinne einer syntakto-semantischen Disambiguierung. Normaldeskriptoren können (evtl. mit Bedeutungsnummern versehene) Einzelwörter sein - darunter fallen hier auch Komposita - oder Syntagmen, wenn es sich um sogenannte Feste Wendungen handelt.

  (2) Teildeskriptoren: Das sind Wörter, die als Ergebnis der Kompositazerlegung und als Teile von Festen Wendungen im Laufe der Dokumentendeskribierung als sinnvolle Einzelelemente erkannt wurden.

  (3) Nicht vollständig vereindeutigte Deskriptoren: Zwar wird bei jedem als lexikalisch mehrdeutig erkannten Wort eine Vereindeutigung versucht; dies gelingt jedoch nicht in allen Fällen. Dann werden aber zumindest von den insgesamt vorhandenen Bedeutungsvarianten eines Wortes einige als unmöglich ausgeschieden. Die verbleibenden Bedeutungsvarianten sind dann Indexierungsergebnisse; allerdings stellen diese verbleibenden Deskriptoren nur ein mögliches Indexierungsergebnis dar. Ihr Zutreffen ist nicht sicher gewährleistet wie im Falle der vereindeutigten Deskriptoren (Normaldeskriptoren). In der Terminologie von CTX werden diese Deskriptoren deshalb als 'möglicherweise richtige' Deskriptoren bezeichnet (in der TELDOK-Realisierung sind es die sog. M-Deskriptoren).

  (4) Komplexe Deskriptoren: Als Ergebnis der bei der automatischen Indexierung vorgenommenen Nominalgruppenanalyse fällt eine Strukturbeschreibung komplexer nominaler Syntagmen an. Erkannt werden verschiedene Formen der Attribuierung (adjektivische, genitivische und präpositionale Attribute) und nominale Anreihungen.

- Lexikalisch-semantische Informationen: es handelt sich dabei um Angaben zur Vereindeutigung syntaktisch und semantisch mehrdeutiger Elemente durch Angabe von Paraphrasen und Beispiele für Verwendungskontexte.

- Fachgebietsbezogene pragmatische Informationen: es wurde eine konzeptuelle Analyse des Fachgebiets, dem die Textgrundlage entstammt, vorgenommen und als Thesaurus in Form begrifflicher Beziehungen fixiert.

### II.6.2.4 Informationsarten

(1) Normaldeskriptoren

Die verschiedenen oben erwähnten Deskriptorklassen bieten einem Benutzer des Informationssystems jeweils unterschiedliche Möglichkeiten, bei einer Suche das Ergebnis zu präzisieren bzw. zu modifizieren. Die im Sinne der Precision besten Ergebnisse wird er dabei durch die Verwendung der Normaldeskriptoren erhalten, da es sich dabei um in ihrer Bedeutung eindeutig bestimmte Zeichenfolgen handelt. Will der Benutzer - etwa zur Erhöhung des Recall - in Kauf nehmen, dass der von ihm verwendete Deskriptor im Verwendungskontext u.U. gewisse Bedeutungsverschiebungen gegenüber der von ihm bei der Suche intendierten Bedeutung erfahren hat, so kann er z.B. Teildeskriptoren verwenden.

Es wäre also wünschenswert, eine logische Trennung dieser in ihren Auswirkungen unterschiedlichen Deskriptorklassen in der Informationsbank zu erreichen.

Hinsichtlich der Benutzbarkeit von Normaldeskriptoren sind jedoch noch einige Überlegungen anzustellen:

- Normaldeskriptoren sind semantisch eindeutige Lexeme. Nun ist die Mehrdeutigkeit sprachlicher Ausdrücke ein Problem, das in normalen Sprachverwendungssituationen nicht gegeben ist. Im allgemeinen ist davon auszugehen, dass die systematische funktionale Mehrdeutigkeit sprachlicher Elemente ein wichtiges Ökonomieprinzip der natürlichen Sprachen darstellt. In einem kontextuell stark restringierten System wie einer Informationsbank, in dem der sonst an der Bedeutungskonstitution maßgeblich beteiligte sprachliche und außersprachliche Kontext fehlt, erweist sich diese systematische Mehrdeutigkeit jedoch als großes praktisches Problem. Dieses Problem wird im CTX-System bei mehrdeutigen Wörtern gelöst durch eine künstliche Erweiterung des Wortschatzes ("normale" Zeichenfolge + Bedeutungsnummer). Diese "kunstsprachlichen" Ausdrücke sind dann zwar im Verwendungskontext "Anfrage an ein Datenbanksystem" eindeutig; es darf jedoch nicht vergessen werden, dass diese Ausdrücke nicht notwendig zum Wortschatz des Sprechers (d.h. Benutzers) gezählt werden dürfen.

- Das Ansetzen von Bedeutungsvarianten ist das Ergebnis einer Problematisierung von Sprache. Für den "unvoreingenommenen" Sprachbenutzer stellen solche Mehrdeutigkeiten aber kein Problem dar. Es ist somit davon auszugehen, dass dem Benutzer die potentielle Mehrdeutigkeit bestimmter Lexeme nicht bewusst ist. Zum sinnvollen Einsatz von vereindeutigten Deskriptoren muss der Rechercheur also bei einem mehrdeutigen Lexem

erst auf dessen Mehrdeutigkeit aufmerksam gemacht werden.

Die Einführung einer Normaldeskriptoren-Ebene in ein Datenbanksystem ist also nicht unproblematisch. Zum sinnvollen Einsatz dieses Retrievalinstruments sollte der Benutzer vom System Unterstützung erhalten, die die beiden eben erwähnten Haupthindernisse aus dem Weg räumt: der Benutzer muss auf Mehrdeutigkeiten aufmerksam gemacht werden, und es müssen ihm die zur Auflösung dieser Mehrdeutigkeit gebildeten künstlichen Ausdrücke mitgeteilt werden.

(2) Komplexe Deskriptoren

TELDOK bietet - wie in IR-Systemen üblich - die Möglichkeit einer Suche mittels boole'scher Verknüpfung; das System eignet sich deshalb zur Realisierung eines Information-Retrieval-Systems, in dem das Prinzip der postkoordinierenden Indexierung angewendet wird. Durch die Analyse der Abhängigkeitsbeziehungen zumindest einer Teilklasse von komplexen sprachlichen Ausdrücken im Lauf der automatischen Indexierung (hier als Komplexe Deskriptoren bezeichnet) ist im CTX-System die Möglichkeit gegeben, zumindest z.T. dem Hauptnachteil strikt postkoordinierender Systeme entgegenzuwirken und themenfremde Zusammenführungen zu vermeiden.

Derartige themenfremde Zusammenführungen treten dann auf, wenn zur Spezifizierung eines komplexen Begriffs bei einer Suchanfrage die Angabe seiner Einzelelemente und einer logischen Verknüpfung dieser Einzelelemente, die das Miteinander-Vorkommen dieser Einzelelemente zum Ausdruck bringt, nicht genügt. Immer wenn die einfachen Einzelelemente auch in anderer bedeutungsvoller Kombination im Text vorkommen können, ist die Möglichkeit themenfremder Zusammenführungen gegeben. In einem strikt postkoordinierenden System besteht die klassische Lösung dieses Problems in der Einführung von Rollenindikatoren, wodurch eine bestimmte Art der strukturellen Abhängigkeit zwischen Elementen einer Suchanfrage zum Suchkriterium gemacht werden kann (vgl. dazu die Indices bei GOLEM). Allgemein gesagt setzt eine solche Problemlösung die Möglichkeit zur Bildung von Relationen über Deskriptoren voraus. Die einzige von TELDOK vorgesehene Möglichkeit zur Darstellung solcher Relationen ist aber schon festgelegt für die semantischen Verknüpfungen. Daher musste zur Ermöglichung komplexer Suchbegriffe in dem Laborsystem JUDO-T ein anderer Weg beschritten werden als in einem strikt postkoordinierenden System: durch Einbringung von präkoordinierten Begriffen, den Komplexen Deskriptoren.

(3) Thesaurus

Im CTX-System liegt ein reich strukturierter Thesaurus des Fachgebiets Datenschutzrecht vor. Es wurde eine Vielzahl von differenzierenden begrifflichen Relationierungen vorgenommen. Diese Verknüpfungsarten unverändert in den Thesaurusteil des TELDOK-Systems zu übernehmen, war wegen der vordefinierten Zahl von Relationen nicht möglich. Deshalb wurde versucht, die im CTX-Thesaurus aufgebauten Verknüpfungen möglichst inhaltsbezogen und ohne größeren Informationsverlust auf die vorgegebenen TELDOK-Relationen abzubilden. Es konnten dabei nur 5 der insgesamt 6 möglichen TELDOK-Relationen für die Abbildung genutzt werden. Die TELDOK-Relation "Homonymie" wurde zur Realisierung des Formaldeskriptor/Normaldeskriptorkonzepts verwendet. Die Sonderbehandlung dieser Relationenart durch TELDOK ermöglicht es, den Benutzer bei der Verwendung eines Formaldeskriptors auf dessen Mehrdeutigkeit aufmerksam zu machen.

Relationale Eigenschaften wie Symmetrie und Transitivität werden in der jetzigen Realisierung der CTX-Anwendung (im Gegensatz zu GOLEM im Fall der Synonymie), nicht ausgewertet; wollte man die sich aus der Transitivität einer Relation ergebenden Begriffsketten in der Datenbank aufbauen, so wäre das nur durch Permutierung aller beteiligten Elemente vor dem Aufbau bzw. der Erweiterung der Datenbank möglich. Da durch die Viele-zu-Eins-Abbildung von Relationen in TELDOK eine große Anzahl von Begriffen jedoch schon ausgiebig vernetzt ist, wurde von dieser Möglichkeit abgesehen, um die Überschaubarkeit der Verknüpfungen nicht zu beeinträchtigen und damit ihre Benutzung nicht zu erschweren.

### II.6.2.5 Abbildung in TELDOK-Strukturen

(1) Thesaurusteil

Die notwendige Trennung von Deskriptorklassen im Thesaurus eines TELDOK-Systems ist grundsätzlich auf zwei Arten realisierbar:

- Markierung der Klassenzugehörigkeit eines Deskriptors durch Kategorien; der Wortlaut des Deskriptors (die Zeichenfolge) bleibt dabei unverändert;

- Markierung durch Kennzeichnung des Deskriptorstrings; die Klassenzugehörigkeit wird integrierter Bestandteil des Deskriptorstrings.

Im CTX-System hat man sich für die zweite Lösung entschieden. Aus experimentellen Gründen sollte eine semantische Relationierung auf der Ebene der Normaldeskriptoren und der M-Deskriptoren durchgeführt werden. Dies wäre, wie man aus dem Kettenstenogramm des Thesaurusteils ersehen kann, bei der ersten Lösung nicht möglich gewesen. Teildeskriptoren und M-Deskriptoren werden in der TELDOK-Version also so dargestellt, dass dem Deskriptorwortlaut zur Kennzeichnung ein 'T' im Falle der Teildeskriptoren und ein 'M' bei den M-Deskriptoren vorausgestellt wird. Diese Kennzeichnung ist vom Deskriptorwortlaut durch ein Leerzeichen getrennt.

Die oben erwähnten Probleme im Zusammenhang mit der Einführung von Normaldeskriptoren wurden dadurch gelöst, dass zu den bereits besprochenen vier Deskriptorklassen (Normaldeskriptoren, Teildeskriptoren, M-Deskriptoren, Komplexe Deskriptoren) noch eine weitere Deskriptorklasse eingeführt wurde: die so genannten Formaldeskriptoren. Diese Deskriptorklasse kann noch am ehesten mit dem "Deskriptor" in einem traditionellen postkoordinierenden System verglichen werden. Es handelt sich um durch keinerlei Zusatz (durch keine Modifikation des Wortlauts) gekennzeichnete Wörter, die prinzipiell mehrere Bedeutungen haben, also Begriffe in einer (Grund-)Form, in der sie dem Sprachbenutzer spontan verfügbar sind. Dieser Deskriptorwortlaut wird in einem Information-Retrieval-System zum Problem, wenn er inhaltlich mehrdeutig ist. Ist dies nicht der Fall, so fällt ein solcher "Formaldeskriptor" in JUDO-T formal mit einem Normaldeskriptor zusammen (Normaldeskriptoren sind definiert als syntaktisch und semantisch eindeutige Begriffe). Von einem Formaldeskriptor im eigentlichen Sinne kann man also erst sprechen, wenn ein mehrdeutiges Wort vorliegt.

Es sei in diesem Zusammenhang daran erinnert, dass eine logische Trennung der (ambigen) Formaldeskriptoren von Normaldeskriptoren in der TELDOK-Version nicht durch Kennzeichnung eines Deskriptorstrings erreicht wird. Vielmehr werden solche Deskriptoren aufgrund der Homo-

nymie-Markierung durch entsprechende Systemreaktionen bei der Recherche für den Benutzer erkennbar. Hält man sich noch einmal vor Augen, dass die Mehrdeutigkeit eines Begriffs einem Sprachbenutzer im Normalfall gar nicht bewusst ist, so ist dies ganz im Sinne des Benutzers. Er hat die Möglichkeit, bei der Suchanfrage Deskriptoren in einer Form zu verwenden, wie sie ihm vom täglichen Sprachgebrauch her vertraut ist (also z.B. das Wort ARM anzugeben). Ist durch die spezifischen Gegebenheiten eines Informationssystems (in diesem Falle z.B., weil im System keine Ausnutzung der Großschreibung möglich ist) der verwendete Begriff mehrdeutig (also nach Definition ein Formaldeskriptor), so wird der Benutzer durch eine Meldung des Systems auf diesen Umstand aufmerksam gemacht. Er erhält als Systemreaktion den Text "BITTE INFODOK KONSULTIEREN". Dieses Beispiel zeigt auch, wie das zweite Problem bei der Benutzung von Normaldeskriptoren gelöst wird. Denn das Wissen, dass ein Ausdruck mehrdeutig ist, reicht bei der Recherche nicht aus; dem Benutzer muss auch noch mitgeteilt werden, welche (kunstsprachlichen) Ausdrücke im System die seiner Absicht entsprechende Bedeutung tragen. Zu diesem Zweck enthält der Dokumententeil der Datenbank analog zur GOLEM-Version eine eigene Dokumentenart (die so genannten Informations-Dokumente oder INFO-Dokumente) mit Metainformationen über Formaldeskriptoren. Diese INFO-Dokumente können vom Benutzer nach Art eines Wörterbuchs während der Recherche konsultiert werden. Es sei noch darauf hingewiesen, dass für Benutzer, die auf eine weitere Precision verzichten oder diese auf andere Weise (z.B. durch boole'sche Verknüpfung) erreichen können, diese Systemreaktion unterdrückt werden kann.

Das Problem der Darstellung Komplexer Deskriptoren wurde gelöst, indem das Prinzip der strikten Postkoordination, bei der nur einfache Begriffe als Deskriptoren zugelassen werden, aufgegeben wurde. Komplexe Deskriptoren sind in der CTX-Konvention Syntagmen, bei denen zwei Einfache Deskriptoren durch ein Kennzeichen der syntaktischen Relation, in der sie im Text stehen, miteinander verknüpft werden. Diese Syntagmen werden vom Informationssystem wie ein (präkombinierter) Deskriptor behandelt.

Bei den Komplexen Deskriptoren macht sich in der TELDOK-Anwendung die Längenbeschränkung bei Deskriptorsätzen des IR-Systems besonders stark bemerkbar. Während bei den anderen Deskriptorklassen Überlängen gelegentlich bei den mehrwortigen Festen Wendungen auftreten, ist die Summe der Zeichen bei Komplexen Deskriptoren häufig größer als die zugelassene Maximallänge von 29 Zeichen. Solche Überlängen müssen in TELDOK-Umsetzung systematisch gekürzt werden (es ist in solchen Fällen jeweils nach dem 13. Zeichen des Einzelelements zu kürzen). Hier zeigt sich zugleich eine Nutzungsmöglichkeit der NATURA-Komponente: Verwendet der Benutzer die Möglichkeit zur Recherche mittels natürlichsprachlicher Problembeschreibung (vgl. Kap. 1.2.8) so wird er von diesem Problem nicht tangiert, da dabei die aus der Analyse der Suchanfrage/Problembeschreibung gewonnenen Deskriptoren (ggf.. durch automatische Längenbeschränkung) auf das in der Informationsbank geltende Format gebracht werden.

Zu allen abgekürzten Deskriptoren wurde ein so genanntes Informationsdokument angelegt, aus dem der volle Deskriptorwortlaut entnommen werden kann. Durch einen TELDOK-Prozedurbefehl kann sich der Benutzer außerdem während der Recherche jederzeit eine Liste aller abgekürzten Deskriptoren, getrennt nach Einfachen und Komplexen Deskriptoren, ausgeben lassen.

(2) Dokumententeil

Der Dokumententeil enthält neben den üblichen Zielinformationen auch die für die Benutzung

von Normaldeskriptoren wichtigen INFO-Dokumente. Die logische Unterscheidung der INFO-Dokumente von anderen Dokumenten besteht in einer differenzierten Kennzeichnung der jeweiligen Dokumentenschlüssel.

Die Dokumentenschlüssel zu den "normalen" Zielinformationen (d.h. zu den mittels Deskriptoren recherchierbaren Texteinheiten) bestehen aus einer mnemotechnischen und eindeutigen Kennzeichnung des Dokuments im Informationssystem, die sich aus einer Textkennzeichnung (z.B. 'BDSG' für 'Bundesdatenschutzgesetz') und einer Nummer des laufenden Dokuments innerhalb eines Texts zusammensetzen. Die einzelnen Teile des Dokumentenschlüssels entstammen den Identifikationsinformationen aus der Ausgabe des Programmsystems zur Dokumentendeskribierung.

Der Dokumentenschlüssel eines INFO-Dokuments ist hingegen ein Wort. Dabei kommen in Frage:

- Formaldeskriptoren: ein Formaldeskriptor ist Dokumentschlüssel für ein Informationsdokument, in dem seine möglichen Bedeutungen (durch Paraphrasen und Verwendungsbeispiele) erklärt sind und diesen unterschiedlichen Bedeutungen die entsprechenden Normaldeskriptoren zugeordnet werden.

- gekürzte Deskriptoren: dabei kommen Deskriptoren aller Klassen in Frage. Das INFO-Dokument enthält in diesen Fällen die Langform des jeweiligen Deskriptors, der Schlüssel ist.

- Normaldeskriptoren: Informationsdokumente können nicht nur semantische Informationen enthalten, wie im Falle der Formaldeskriptoren, sondern auch fachgebietsbezogene juristische Informationen wie Begriffsdefinitionen und Ouellenangaben. Solche Informationen können auch den Paraphrasen von Formaldeskriptoren beigegeben sein.

Die Verwendung von deskriptorfähigen Elementen als Dokumentenschlüssel erlaubt den problemlosen Zugriff auf Informationsdokumente zu jedem Zeitpunkt während der Recherche.

### II.6.3 Index- und Konkordanzkomponente

Neben dem Anschluss an on-line-orientierte Information-Retrieval-Systeme ist auch der Aufbau einer Katalogkomponente ein wichtiges Desiderat, da sich daraus gegenwärtig weitere Anwendungen ableiten lassen. Es erschien daher sinnvoll, das CTX-System um einen derartigen Baustein anzureichern.

Unter der Index- und Konkordanzkomponente wird die Erstellung von Wort-Indices, Konkordanzen auf Satz- bzw. Dokumentebene sowie die Aufbereitung von Texten nach verschiedenen Gesichtspunkten wie Häufigkeitsregister, rückläufige Wortlisten u.a.m. verstanden.

Da die Ausgabe von Indices und Konkordanzen - ebenso wie von Analyse-Ergebnissen - auch schon bei Texten mit geringem Umfang viele Seiten beansprucht, ist (in Teilbereichen) auf die Ausgabe über Papier verzichtet worden. Stattdessen erfolgt die Ausgabe auf Microfiche, wobei sich (bei 48-facher Verkleinerung) 270 Seiten Computerausdruck (1 Seite entspricht dabei ca. 1,5

DIN A4 Seiten) auf einen Fiche der Größe DIN A6 komprimieren lassen.

Die so erstellten Microfichedaten sind einerseits in der Entwicklungs- und Experimentierphase des Projekts JUDO-DS heim Test der Ergebnisse sowohl der linguistischen Analyse als auch der verschiedenen Schnittstellen eine wertvolle Hilfe.

Andererseits bietet die Katalogkomponente einem künftigen Benutzer des CTX-Systems auch ohne Terminal die Möglichkeit, eine Art 'Off-line-Retrieval' durchzuführen, natürlich verbunden mit Einschränkungen hinsichtlich der Verknüpfung von Deskriptoren. Grundlage dieses Retrievals sind dabei Indices oder Konkordanzen eines oder mehrerer Texte. (Am Rande sei hier nur die Möglichkeit einer Variante eines Dokument-Informations-Systems mit Microfiche-Komponente erwähnt, das in etwa mit den auf Microfiche vorliegenden Steuerentscheidungstexten bei der DATEV zu vergleichen wäre.)

Zum jetzigen Zeitpunkt sind einige Programme realisiert, die aus einem Text, basierend auf den Ergebnissen der automatischen Sprachanalyse Wort-Indices und Satz-Konkordanzen erstellen. Im Projekt JUDO-DS dienten diese Programme in erster Linie der Test-Unterstützung und erst in einer Art Nebeneffekt dazu, weitere Auswertungs- und Informationsmöglichkeiten aufzuzeigen und zu erproben.

## II.7 Exkurs: Vergleich der Indexierungsergebnisse von CTX und JURIS/PASSAT

Durch einen exemplarischen Vergleich der Dokumentdeskribierung von CTX mit den Ergebnissen einer GOLEM/PASSAT-Anwendung bei JURIS (Freitextbereich) wird versucht aufzuzeigen, welche Fortschritte und Verbesserungen durch den Einsatz einer umfangreicheren linguistischen Analyse gegenüber alternativen Verfahren der automatischen Indexierung erreicht werden. Der Vergleich wurde 1980 durchgeführt.

Von 1975 bis 1980 bestand im Rahmen des Forschungsprojekts ein Anschluss an JURIS, so dass die technischen Voraussetzungen für den Vergleich mit GOLEM/PASSAT in der JURIS-Anwendung vorlagen (im Gegensatz zu anderen vergleichbaren Verfahren wie STAIRS/TLS). Als Vergleichsgrundlage wurde das Bundesdatenschutzgesetz (BDSG) verwendet. Es liegt auch im Rahmen des JURIS-Systems bearbeitet und (über PASSAT) deskribiert vor.

Es muss bei dieser Gegenüberstellung darauf hingewiesen werden, dass die Vergleichsdaten von JURIS aus einer konkreten, allgemeinen PASSAT-Anwendung gewonnen wurden, während die Modellanwendung des CTX-Systems aus der konzentrierten Behandlung von Datenschutztexten durchaus Vorteile gezogen haben mag. Insofern sind die Vergleichsergebnisse zu relativieren. Andererseits musste im Interesse der Beschreibung der Projektergebnisse eine anschauliche Vergleichsbasis gewählt werden. Für die Untersuchung steht insofern das Verfahren PASSAT stellvertretend für zeichenkettenorientierte Systeme.

### II.7.1 Dokumentdeskribierung in beiden Systemen

Die ausführliche Darstellung des Vorgehens bei der Dokumentdeskribierung im CTX-System geht aus den vorausgehenden Kapiteln hervor. Zum besseren Verständnis des folgenden seien hier nochmals die einzelnen Deskriptorarten aufgelistet:

- "Einfache Deskriptoren" umfassen hier die deskriptorfähigen "lemmatisierten", d.h. auf Grundformen reduzierten Textwortformen und die durch die semantische Analyse zusammengeführten Begriffe, die sog. Festen Wendungen (z.B. DATENUEBERMITTLUNG, JURISTISCHE PERSON).

- "Komplexe Deskriptoren" auf der Grundlage syntaktischer Relationen, geben ausgewählte, durch die linguistische Analyse gewonnene Oberflächenstrukturen des Textes wieder, wobei als äußeres Kennzeichen ein Relator (G,P,K) vergeben ist, z.B. "Übermittlung von Daten" -- UEBERMITTLUNG G DATUM.

Bei GOLEM/PASSAT in der JURIS-Anwendung werden zur textuellen Erschließung folgende Deskriptoren vergeben:

- dokumentarische Angaben im "Überbau" eines jeden Dokuments in Form von gebundenen Deskriptoren;
- intellektuell zugeteilte Schlagwörter;
- automatisch lemmatisierte Textwortformen.

Jedes Dokument von JURIS wird bei seiner Erfassung mit einem dokumentarischen Überbau versehen. Dieser enthält bei Normen (die hier verglichen werden.) in über 40 Rubriken Angaben, z.B. über Art der Norm (Gesetz), Normgeber, Inkrafttreten sowie Gültigkeitsdauer und -bereich der Norm. Als weitere Erschließungsmittel dienen intellektuell zugeteilte Schlagwörter, die im Dokumenttext selbst in der Regel nicht vorkommen, jedoch die aus dem Text automatisch erschlossenen Wörter um relevante Begriffe ergänzen (Beispiel: ein Dokument mit dem Textwortlaut "zur Verschwiegenheit verpflichtete Person" erhält zusätzlich zu Wörtern wie VERSCHWIEGENHEIT, VERPFLICHTET und PERSON das Schlagwort VERSCHWIEGENHEITSPFLICHT). Schließlich werden die durch automatische Indexierung über PASSAT gewonnenen Deskriptoren im Retrieval zur Verfügung gestellt.

### II.7.2 Automatische Indexierung durch PASSAT

PASSAT soll eine (kontextfreie, d.h. zeichenkettenorientierte) morphologische Reduktion der Textwortformen auf Grundformen leisten. Als Basis dazu dient die intellektuell erstellte und gepflegte "Vergleichswortliste" (VWL), die die sog. "Stammwörter" enthält, versehen mit weiteren Kodierungen (s.u.). Unter einem Stammwort wird "... diejenige Zeichenfolge eines Wortes, die in allen flektierten Formen gleich bleibt" verstanden (SIEMENS 1980, S.5). Stammwörter sind:

- Nominativ Singular von Substantiven
- Umlautplurale (z.B. FAELLE)
- Komposita, die nicht sinnvoll zertrennt werden können
- mehrwortige Begriffe
- bestimmte Fremdwörter, deren Stamm sich hei der Flexion ändert, in verstümmelter Form (z.B. MATRI mit den Endungen -X, -CES, -ZES, -ZEN)
- Infinitiv ohne Endung bei Verben (z.B. SAG)
- unregelmäßige Stammformen von Verben
- Partizipien mit Präfix oder Infix
- steigerungsfähige Gerundformen

- Positivformen von Adjektiven
- unregelmäßige Steigerungsformen von Adjektiven
- Pronomen ohne Flexionsendung
- Eigennamen
- Zahlen
- Präpositionen, Konjunktionen, bestimmte Artikel.

In den jedem Eintrag beigefügten Markierungen wird auf je eine der drei folgenden Listen verwiesen: die Wortklassenliste, die Endungsliste und die Bindungsliste. In der Wortklassenliste ist angegeben, zu welcher Wortart das Stammwort gehört. In dieser Information ist auch enthalten, ob das Wort als deskriptorfähig oder als Stoppwort eingestuft werden soll. Der Hinweis auf die Endungsliste führt zu einem Eintrag, in dem alle möglichen Endungen, die das Stammwort annehmen kann, erfasst sind. Jeder Eintrag der Endungsliste ist so aufgebaut, dass durch Anfügen des ersten Elements eines solchen Eintrags an das Stammwort gerade die als Deskriptor zu vergebende Form entsteht. In der Bindungsliste ist schließlich angegeben, ob und mit welchen Fugenmorphemen ein Stammwort Teil eines Kompositums sein kann.

Als Deskriptoren werden dann vergeben:

- deskriptorfähige Stammwörter nach Anfügen der ersten Endung des entsprechenden Eintrags in der Endungsliste;
- Komposita;
- alle identifizierten Bestandteile der Komposita und alle nach Vergleich mit der Bindungsliste erlaubten Verknüpfungen dieser Bestandteile;
- Ausdrücke, die in der VWL als mehrwortige Begriffe gekennzeichnet sind, werden zusätzlich zu ihren Einzelbestandteilen vergeben.

### II.7.3 Vergleich der unterschiedlichen Deskriptorarten

Der von JURIS für alle Dokumente erstellte Überbau stellt ein übliches Klassifikationsmittel für derartige Dokumente dar. Er strukturiert v.a. dokumentbezogene allgemeine Angaben. Daneben ist es eine Frage der Ökonomie, der Benutzerakzeptanz und des Benutzerverhaltens, ob zusätzliche intellektuelle Schlagwörter zu jedem neu erfassten Dokument einzeln vergeben werden oder ob solche Beziehungen für alle Dokumente über einen Thesaurus erschließbar sind.

Die im CTX-System über die prinzipiellen Möglichkeiten von JURIS (GOLEM/PASSAT) unmittelbar hinausgehenden Texterschließungsmittel betreffen die syntaktische und semantische Vereindeutigung mehrdeutiger Textwörter sowie den Aufbau von syntaktischen Relationen, d.h. die Präkoordination von Deskriptoren. Hierzu bietet GOLEM mit dem Subsystem PASSAT nur eine partielle Lösung (Grundformenermittlung) an.
Es soll an dieser Stelle nicht diskutiert werden, welchen praktischen Nutzen eine verfeinerte Texterschließung i.S. einer Mehrwortermittlung bzw. einer semantischen Disambiguierung haben kann. Die Vermutung, dass das Problem der Mehrdeutigkeiten von Deskriptoren durch die Verknüpfungslogik beim Retrieval weitgehend ausgeschlossen werden kann, ist bislang - eben aus dem Grund, dass keine entsprechenden Verfahren vorliegen - nicht bestätigt worden. Die Korrektheit und Vollständigkeit der syntaktischen Relationen ist zudem weitgehend von der erreichten Analysequalität abhängig; es ist jedoch einsichtig, dass diese Relationen einen wesentlichen

Beitrag zur Erhöhung der Precision leisten können. Die in Systemen wie STAIRS oder DIRS/ GRIPS realisierten Funktionen der "Adjacency" bestätigen eher die Notwendigkeit solcher Präzisierungsfunktionen.

### II.7.4 Vergleich der Einfachen Deskriptoren

Ein Vergleich PASSAT - CTX ist nur auf der Ebene der sog. "Einfachen Deskriptoren" sinnvoll, da die Grundlage für einen Vergleich der Komplexen Deskriptoren fehlt.

(1) Erstellung der Vergleichsbasis anhand des Bundesdatenschutzgesetzes (BDSG)

Als Vergleichsbasis wurde eine Liste erstellt, bestehend aus allen Wörtern bzw. mehrwortigen Begriffen, die durch CTX als einfache Deskriptoren zum BDSG vergehen wurden sowie eine ebensolche Liste mit Begriffen, die von JURIS den Dokumenten zum BDSG intellektuell sowie als durch PASSAT gewonnene Deskriptoren zugeteilt wurden. Die Einteilung des BDSG in Einzeldokumente stimmt bei beiden Systemen nicht vollständig überein. Nicht berücksichtigt werden bei dieser Untersuchung die Unterschiede, die sich aus der daraus resultierenden Zuordnung der Deskriptoren zu den einzelnen Dokumenten eventuell ergeben.

Als deskriptorfähig gelten in beiden Systemen alle Wortformen, die von Verben, Substantiven und Adjektiven abgeleitet sind, sowie Eigennamen. Numeralia werden in beiden Systemen unterschiedlich behandelt: Bei JURIS stehen sie nur als Paragraphenzählung (z.B. in der Wendung 'BDSG § 1') zur Verfügung. Bei CTX werden sie bislang sowohl als Einzeldeskriptoren vergeben wie auch zu zusammengesetzten Ausdrücken verbunden (z.B. § 2, Abs. 4). In den folgenden Abschnitten sind die Numeralia nicht mehr berücksichtigt (auch nicht in den statistischen Werten).

| | |
|---|---|
| Insgesamt wurden vergeben | 1573 Deskriptoren |
| davon von JURIS | 1277 " |
| von CTX | 1055 " |
| beiden Systemen waren im Wortlaut gemeinsam | 759 " |

Die gemeinsamen Deskriptoren teilen sich folgendermaßen auf:

| | |
|---|---|
| Verbformen | 192 |
| Adjektive | 119 |
| Substantive | 439 |
| mehrwortige Begriffe | 6 |
| sonstige | 3 |

Unter den Verbformen finden sich neben Infinitivformen die von CTX als attributiv gebraucht erkannten und deshalb zusätzlich zu den Infinitiven vergebenen Partizipformen. Bei den Substantiven treten nur Nominativ-Singular-Formen auf. Die mehrwortigen Begriffe werden vorerst nur zahlenmäßig erfasst; inhaltlich behandelt werden sie gesondert in Kap. II.7.5.
Die beiden Ergebnissen gemeinsamen Deskriptoren wurden nicht weiter auf ihre richtige Lemmatisierung überprüft, da in einer Reihe früherer Tests festgestellt worden war, dass die Reduktion auf Grundformen von CTX nahezu vollständig richtig erfolgt.

(2) Nur von CTX vergebene Deskriptoren

Aus den Zahlen im vorhergehenden Abschnitt ergibt sich, dass von den von CTX vergebenen Deskriptoren 296 nicht von JURIS vergeben wurden. Diese lassen sich zahlenmäßig folgendermaßen aufschlüsseln:

| | |
|---|---|
| Verben | 96 |
| Substantive | 14 |
| Adjektive | 19 |
| Komposita | 22 |
| mehrwortige Begriffe | 142 |
| Abkürzungen | 3 |

Diese Wortformen fehlen bei JURIS aus folgenden Gründen:

- PASSAT leistet nur eine eingeschränkte Lemmatisierung von Verben auf ihre Grundform; so wird z.B. bei durch Ablaut gebildeten Stammformen von starken Verben sowie bei durch Prä- oder Infigierung gebildeten Partizipien nur eine evtl. vorhandene Endung abgetrennt, diese Formen werden aber nicht auf den Infinitiv zurückgeführt. Z.B. wird die Textwortform BESCHLOSSEN von CTX auf BESCHLIESSEN und bei PASSAT auf BESCHLOSS zurückgeführt. Ein anderes Beispiel: Die Textform ANZUWENDEN bleibt bei JURIS/PASSAT unverändert, während sie bei CTX auf ANWENDEN lemmatisiert wird. Außerdem werden bei CTX im Gegensatz zu JURIS/PASSAT die im Text diskontinuierlich auftretenden Verbformen zusammengeführt (z.B. wird FORDERT ... AUF zu AUFFORDERN).
- Die nur bei CTX vorhandenen Substantive (an dieser Stelle ohne Berücksichtigung der Komposita) sind von JURIS/PASSAT falsch oder gar nicht lemmatisiert: so wird z.B. die Textwortform ZEUGE zu ZEUG oder ZEUGEN oder es bleibt die Textwortform NAMEN unverändert. Die adjektivisch deklinierten Substantive werden von PASSAT anders behandelt als von CTX: PASSAT ordnet z.B. den Textwortformen BETROFFENE, BETROFFEN, BETROFFENER nur die Form BETROFFENE zu, wohingegen sich bei CTX auch noch die Form BETROFFENER findet.
- Die Adjektive sind teils unterschiedlich lemmatisiert (z.B. ANDERE von PASSAT im Vergleich zu ANDER bei CTX), teils sind sie von PASSAT zerlegt, ohne dass das ursprüngliche Adjektiv selbst als Deskriptor erscheint (z.B. für DIENSTRECHTLICH ist nur DIENST und RECHTLICH vorhanden). Es sind aber auch nicht immer alle Zerlegungsergebnisse von Adjektiven in der Deskriptorliste vorhanden (ein Zerlegungsergebnis wird nicht aufgenommen, wenn es als Stoppwort klassifiziert wird). Unter den Adjektiven, die bei JURIS völlig fehlen, sind hauptsächlich solche wie BESONDER, SOLCH, BISHERIG und Zahladjektive wie ZWEIT, DRITT, etc.
- Bei 21 der fraglichen Komposita finden sich in der JURIS-Deskriptorliste nur die zugehörigen Simplicia (z.B. bei GERICHTSVERFAHREN nur GERICHT und VERFAHREN); zwei Komposita sind falsch lemmatisiert (EINSICHTNAHME zu EINSICHTNAHM und KENNTNISNAHME zu KENNTNISNAHM), ohne dass die richtige Grundform erfasst wurde.

(3) Nur von JURIS/PASSAT vergebene Deskriptoren

Die Zahl der nur von JURIS vergebenen Deskriptoren setzt sich folgendermaßen zusammen:

| | |
|---|---|
| zusätzliche Schlagwörter (JURIS) | 81 |
| Verbformen | 130 |
| Substantivformen | 199 |
| Adjektivformen | 31 |
| nicht deskriptorfähige Einheiten | 5 |
| falsche Grundformen | 52 |
| mehrwortige Begriffe | 28 |

Die Unterschiede zu CTX ergeben sich aus folgenden Gründen:

- Bei JURIS werden zusätzliche Schlagwörter intellektuell als Deskriptoren vergeben. Diese Schlagwörter lassen sich in zwei Klassen unterteilen: die einen bilden eine inhaltliche Zusammenfassung eines Dokuments zu im Text nicht vorhandenen Begriffen (wie z.B. ANHOERUNGSRECHT), die anderen sind morphologische Ableitungen wie VERWEIGERUNG von VERWEIGERN. Beide Arten von Begriffen werden von CTX nicht als einfache Deskriptoren vergeben, es wird jedoch versucht, sie über Thesaurusrelationen zu erfassen; dies ist bei Derivationen wie VERWEIGERUNG und VERWEIGERN die Regel.
- Wie schon erwähnt, wird von PASSAT bei Verben nur eine Endungsanalyse durchgeführt, daher ergeben sich bei den Verbformen folgende Zahlen:

  | | |
  |---|---|
  | Infinitivformen | 49 |
  | Infinitiv mit infigiertem ZU | 23 |
  | Partizip I | 1 |
  | Partizip ll | 36 |
  | sonstige Verbformen (Präteritum, etc.). | 21 |

  Die 49 Infinitivformen sind durch PASSAT alle falsch lemmatisiert, dabei treten hauptsächlich folgende Fehlerquellen auf: Wortformen, die nicht zur Wortklasse Verb gehören, werden auf Verben zurückgeführt (z.B. SCHON (ADV) auf SCHONEN, WEIL (Konj.) auf WEILEN), ohne dass jedoch die Wortform selbst in der Deskriptorliste auftritt, da sie als Stoppwort klassifiziert wird. Verbformen, die im Text diskontinuierlich auftreten, werden naturgemäß nicht zusammengeführt (z.B. anstelle von AUFFORDERN ist nur FORDERN vergeben). Deskriptorfähige Wortformen mit Endungen, die auch als Flexionsendungen von Verben möglich sind (z.B. T bei RECHT), führen dazu, dass diese Wortformen fälschlicherweise auch auf Infinitive abgebildet werden. In diesen Fällen war aber immer zusätzlich der richtige Deskriptorwortlaut vorhanden.

  Im System CTX wird bei den Partizipformen zwischen den attributiv und prädikativ gebrauchten Partizipien unterschieden; so sind alle nur bei JURIS auftretenden Partizipien im Text prädikativ gebraucht und werden deshalb bei CTX auf den Infinitiv zurückgeführt. Unter sonstige Verbformen fallen ablautende Stammformen starker Verben, die bei CTX über das syntaktische Lexikon auf ihre Grundform reduziert werden (z.B. BETRAEGT auf BETRAGEN).

- Bei den Substantivformen handelt es sich neben den falsch reduzierten Formen (z.B. RECHTS (Genitiv Singular) bleibt RECHTS), den nicht lemmatisierten Formen (z.B. NAMEN) und den intellektuell vergebenen Schlagwörtern hauptsächlich um Komposita und ihre Zerlegungsergebnisse. Im System CTX wird - wie erwähnt - die von der Textanalyse angebotene automatische Kompositumzerlegung nicht unkontrolliert benutzt: die automatischen Zerlegungsergebnisse werden intellektuell kontrolliert und dann als Relationen in den CTX-Thesaurus übernommen (vgl. Kap. II.3.6). Somit treten also die Zerlegungselemente bei CTX nicht als einfache Deskriptoren auf. Die algorithmische Zerlegung der Komposita durch PASSAT führt zu den hinlänglich bekannten Fehlern. Zwar könnten durch Aufnahme ganzer Wortstämme sinnlose Zerlegungen vermieden werden (z.B. die Zerlegung von BESTIMMUNGEN in BEST, IMMUN und GEN durch Aufnahme von BESTIMMUNG in die VWL); tritt dieser Wortstamm jedoch selbst wieder als Teil eines nicht in der VWL enthaltenen Kompositums (z.B. BEGRIFFSBESTIMMUNGEN) auf, so werden als Deskriptoren wieder alle möglichen Zerlegungen samt ihren Verknüpfungen angeboten (in obigem Beispiel: BEGRIFF, BEST, IMMUN, GEN, BESTIMMUNG, BEGRIFFSBESTIMMUNGEN). Weitere Beispiele dafür sind ZWEI, FELS, FALL und ZWEIFEL für ZWEIFELSFALL oder UR und KUNDE für URKUNDE (Zweifel, Fall und Urkunde sind in der VWL enthalten).

- Die meisten Adjektive, die von JURIS, nicht aber von CTX als Einfache Deskriptoren vergeben wurden, resultieren aus der Zerlegung von zusammengesetzten Wortformen. Dabei treten neben morphologisch und semantisch sinnvollen Zerlegungsergebnissen wie BEHÖRDLICH von INNERBEHÖRDLICH oder ZIVIL von ZIVILPROZESSORDNUNG, auch zwar morphologisch korrekte, aber semantisch unzutreffende Teilelemente auf wie MAESSIG von SATZUNGSMAESSIG.

  Dazu kommen Adjektivformen, die aus falschen Zerlegungen stammen, wie z.B. IMMUN von BESTIMMUNG. Abgetrennte Verbzusätze wie (FEST von LEGT ... FEST) können durch PASSAT nicht mit ihrem zugehörigen Simplex zusammengeführt werden. Die restlichen Unterschiede bei Adjektiven ergeben sich aus der Zurückführung auf verschiedene Grundformen (z.B. ANDERE bei PASSAT und ANDER bei CTX) und der unterschiedlichen Behandlung indefiniter Substantive (z.B. ANGEHOERIG (PASSAT)) im Vergleich zu ANGEHOERIGER (CTX))

- Die Erzeugung falscher Grundformen hat folgende Ursachen: Textwortformen werden auf alle Grundformen abgebildet, die potentiell zutreffen können. Kontextfrei können aber diese Formen nicht von der richtigen Form unterschieden werden (z.B. WEIL (Konj.) wird lemmatisiert auf WEILEN (Verb) oder RECHT auf RECHEN). Diese Fälle wurden schon unter den obigen Rubriken behandelt. Weiter wurden Textwortformen auf Grundformen reduziert, die im deutschen Sprachgebrauch nicht üblich sind (z.B. ABSCHIRMDIENEN von ABSCHIRMDIENST). Die richtige Grundform war in nahezu allen diesen Fällen zusätzlich vorhanden. Nicht aufgelistet sind in dieser Rubrik falsche Zerlegungsergebnisse von zusammengesetzten Textwortformen; diese wurden schon bei der Untersuchung der Wortformen aus den Hauptwortklassen Substantiv, Verb und Adjektiv behandelt.

### II.7.5 Vergleich der mehrwortigen Begriffe

Von beiden Systemen werden mehrwortige Begriffe (bei CTX die sog. Festen Wendungen) als freie Deskriptoren vergeben. Während CTX auch diskontinuierlich auftretende Einheiten zusammenführen kann, werden mehrwortige Begriffe von PASSAT nur dann automatisch als Deskriptoren erstellt, wenn alle Teilelemente (auch flektiert) in unmittelbarer Aufeinanderfolge im Text vorhanden sind. Im folgenden sei ein zahlenmäßiger Überblick über die von CTX und JURIS zum BDSG vergebenen mehrwortigen Begriffe aufgeführt:

| | |
|---|---|
| mehrwortige Begriffe insgesamt | 175 |
| bei CTX und JURIS gemeinsam | 6 |
| nur bei JURIS | 28 |
| nur bei CTX | 141 |

- Bei den nur bei JURIS vorhandenen mehrwortigen Begriffen handelt es sich - bis auf eine Wortfolge, die bei CTX in lemmatisierter Form (ALLGEMEINE VERWALTUNGS-VORSCHRIFT(EN)) vorliegt - ausschließlich um mehrwortige Abschnitts- und Paragraphenüberschriften, die intellektuell (zusätzlich zu PASSAT) vergeben wurden (z.B. AUFGABE UND GEGENSTAND DES DATENSCHUTZES (§ 1) oder WEITERGELTENDE VORSCHRIFTEN (§ 45)). Überschriften wurden von CTX nicht per se aufgenommen; die darin enthaltenen inhaltlichen Beziehungen sind jedoch in allen Fällen über die syntaktischen Relationen als Komplexe Deskriptoren für das Retrieval zur Verfügung gestellt (in obigem Beispiel also: AUFGABE G (Genitiv) DATENSCHUTZ, GEGENSTAND G DATENSCHUTZ, AUFGABE K (Konjunktion) DATENSCHUTZ, WEITERGELTENDE VORSCHRIFT (Adjektivrelation)).

- Die nur bei CTX registrierten mehrwortigen Begriffe sind nach ihrer Verarbeitungsweise bei der syntaktischen Analyse in zwei Klassen einzuteilen:

  In der einen Klasse sind die Festen Syntagmen als unflektierbare kontinuierliche Folgen von Textwörtern enthalten (z.B. AUF GRUND VON); diese Wortfolgen sind als eigene Einträge in das syntaktische Wörterbuch aufgenommen und werden bereits bei der Wörterbuchsuche zusammengeführt. Von dieser Art waren 34 der im CTX-System identifizierten mehrwortigen Begriffe.

  Bei der zweiten Klasse handelt es sich um die eigentlichen Festen Wendungen. Diese Klasse ist in ihrer Zusammensetzung weit weniger homogen; als gemeinsames Merkmal aller ihr zugeordneten Begriffe gilt jedoch, dass die Zusammenführung der einzelnen Textwortformen zu einem Begriff erst nach erfolgreicher syntaktischer Bearbeitung durch die semantische Komponente des Sprachanalyseverfahrens geleistet wird. Diese Festen Wendungen können im Gegensatz zu den oben angeführten Festen Syntagmen im Text auch diskontinuierlich auftreten. Unter diese Festen Wendungen fallen syntaktische und semantische Sequenzen, deren Bestandteile für sich genommen keinen Sinn ergeben oder die bei einer Zertrennung ihre Bedeutung ändern (z.B. AUSSER KRAFT TRETEN oder JURISTISCHE PERSON). Ausgehend vom Fachgebiet Datenschutz wurde sodann der Begriff "Feste Wendung" im CTX-System ausgedehnt auf solche Wortfolgen, deren Bedeutung sich zwar aus der Bedeutung der Einzelbestandteile ergibt, die jedoch in eben ihrer charakteristischen Anordnung häufig im Text belegt und insbesondere dem "Fachmann" in dieser Anordnung geläufig sind (z.B. BUNDESBEAUFTRAGTER FUER DEN

DATENSCHUTZ).

### II.7.6 Vergleichsergebnisse

Die untersuchten Daten bestätigen die Annahme, dass die automatische Reduktion der Textwortformen auf richtige Grundformen von CTX nahezu vollständig geleistet wird. Ebenso kann von einer vollständigen syntaktischen Disambiguierung für die Deskriptorwortklassen Substantiv, Adjektiv und Verb ausgegangen werden. Auch für PASSAT gilt das Problem der Lemmatisierung im wesentlichen als gelöst, wenn man von der fehlenden Zusammenführung verschiedener Verbstämme zu einer gemeinsamen Grundform und von einer gewissen Deskriptorredundanz absieht. In der Mehrzahl der Fälle existiert bei JURIS/PASSAT zu einer falsch reduzierten Textwortform auch die richtige Lösung. Dies kann aber naturgemäß in manchen Fällen Probleme im Hinblick auf Recall und Precision nach sich ziehen. Eine Auflösung von syntaktischen Homographen kann von einem kontextfreien System wie PASSAT nicht erwartet werden.

Vergleicht man den Aufwand, der bei beiden Systemen zur Erzielung dieser Ergebnisse nötig ist, liegen mit Einschränkung auf der lexikalischen Seite keine großen Unterschiede vor. Die für die Wörterbuchsuche und die syntaktische Disambiguierung bei CTX notwendigen Informationen im syntaktischen Wörterbuch umfassen Stammeintrag, Lemmanamen, Wortklassenkodierung und Angaben zu den Flexionsendungen sowie Kodierungen für abtrennbare Verbzusätze. Sieht man also von der Zuordnung des Lemmanamens zur Stammform und von den Markierungen zu den Verbzusätzen ab, so entsprechen diese Angaben im Großen und Ganzen denjenigen der Vergleichswortliste bei PASSAT.

Darüber hinausgehende syntaktische Angaben sind in diesem Zusammenhang nicht zu berücksichtigen, da sie erst zu einer weiterreichenden Analyse benötigt werden, durch die es bei CTX möglich wird, weitere Textspezifika durch die syntaktischen Relationen und die Disambiguierung semantischer Mehrdeutigkeiten für das Retrieval zu erschließen. Die letztgenannten Texterschließungsmittel konnten aber hier nicht Gegenstand des Vergleichs sein.

Die Daten, die die Vergleichsgrundlage bildeten, können - wegen der wesentlich verschiedenen Grundlagen und Ziele der beiden Systeme - keinen absoluten Beweis antreten für die Überlegenheit des einen über das andere System. Es wird auch in Zukunft Anwendungsfälle geben, in denen weniger aufwendige Systeme wie PASSAT ausreichen werden. Andererseits lassen sich auch in kommerzielle Indexierungs- und Retrievalsysteme Teilkomponenten (wie eine verbesserte Wörterbuchsuche, exakte Verfahren zur Kompositazerlegung, Thesauri) einbringen, die schon entscheidende Fortschritte gegenüber dem jetzigen Stand der Technik bringen, ohne dass ein Syntaxparser u.a.m. zum Einsatz kommen müssen.

## GLOSSAR

CUT-OFF-WERT:
Statistisch orientiertes Vorgehen bei CTX, das Bedeutungsvarianten in Abhängigkeit von einem numerischen Cut-Off-Wert selektiert.

DESKRIBIERUNG:
(Automatische) Zuteilung von Stichwörtern zur Identifizierung von Dokumenten.

DESKRIPTOREN:

(1) Einfache Deskriptoren: Einwortige Suchbegriffe, die Wörter der Wortklassen Substantiv, Adjektiv und Verb umfassen

(2) Komplexe Deskriptoren: Auf syntaktischer Ebene verbundene Einfache Deskriptoren. Es gibt verschiedene Typen von Komplexen Deskriptoren:

- (a) Adjektiv-Substantiv-Verbindung
  Beispiel: EINSTIMMIGE ENTSCHEIDUNG
- (b) G-Relation (Genitiv-Relation)
  Beispiel: UNTERNEHMEN G PRESSE (G = Genitiv, hier: "der")
- (c) K-Relation (Anreihungs-Relation)
  Beispiel: DISKUSSION K FOERDERUNG (K=Konjunktion; hier: "und")
- (d) P-Relation (Präpositional-Relation)
  Beispiel: ZUGANG P INFORMATION (P=Präposition; hier: "zu")
- (e) AKK-VRB-Relation (Akkusativ-Verb-Relation)
  Beispiel: DATEN VERARBEITEN
- (f) MOD-VRB-Relation (Modalverb-Verb-Relation)
  Beispiel: VERARBEITEN KÖNNEN

INFORMATION-RETRIEVAL:

(1) Verfahren, in Datenbeständen Suchvorgänge durchzuführen

(2) Methoden der Speicherung bzw. Darstellung der ausgewerteten Elemente in Datenbanken

(3) (Automatisierung der) Inhaltserschließung

LEMMA:
Grammatische Grundform (s.a. LEMMATISIERUNG)
Beispiel: HAUS, SCHÖN, FAHREN

LEMMATISIERUNG:
Reduktion von Wortformen auf ihre Grundform =Lemma)
Beispiele:

| Morphologische Erscheinung | Textwortform | Grundform |
|---|---|---|
| Umlaut bei Substantivpluralbildung: | Vorzüge | Vorzug |
| Stammveränderung bei starken Verben: | brachte | bringen |
| Affixabtrennung: | trifft...zu | zutreffen |
| Vollständige flexionsbedingte Graphemfolgenverschiedenheit: | besser | gut |

MORPHOLOGISCH:

Die Form von Wörtern betreffend

MORPHO-SYNTAKTISCH:
Äußere Form und syntaktische Funktion von Wörtern betreffend

OBERFLÄCHENSYNTAKTISCHE ERSCHEINUNGSFORM:
Form eines Ausdrucks, wie er in der konkreten Äußerung erscheint;
der Analyse unmittelbar zugängliche Ebene der syntaktischen Struktur

PRECISION:
Das Verhältnis der Anzahl der gefundenen relevanten Dokumente zu der Anzahl der Dokumente, die insgesamt gefunden wurden

RECALL:
Das Verhältnis der Anzahl der gefundenen relevanten Dokumente, zu der Anzahl der vorhandenen relevanten Dokumente

SEMANTISCH:
Die Bedeutung von Wörtern betreffend

SEMANTISCHE DISAMBIGUIERUNG: Vereindeutigung von mehrdeutigen Begriffen

THESAURUS:
Wörterbuch, in dem die Einträge bedeutungsmäßig miteinander verknüpft sind

TOKEN:
Laufende Wortformen

TRUNCATION/TRUNKIERUNG (in Zeichen:$)
Mittel der Zusammenführung von Zeichenketten mit gemeinsamen Teilstrings (Teilketten)
Beispiel: TISCH, TISCHTENNIS, usw. werden gemeinsam deskribiert durch; TISCH$

**ABKÜRZUNGSVERZEICHNIS**

A: Adjektiv

A: ADJ-Relation

ABK: Bezeichnung für die Relation "Abkürzung"

ADJ: Bezeichnung für die Wortklasse "Adjektiv"

ADJ-Relation: Zur Kennzeichnung einer Beziehung zwischen ADJektiv-Attribut und einem Substantiv in einem Komplexen Deskriptor

ADV: ADVerb

ASS: ASSoziationsrelation

BDSG: Gesetz zum Schutz vor Missbrauch personenbezogener Daten bei der Datenverarbeitung (BundesDatenSchutzGesetz), vom 27.1.77, BGB1. I S.201

BMFT: BundesMinisterium für Forschung und Technologie, Bonn

BMJ: Bundes-Ministerium der Justiz, Bonn

BTX: BildschirmTeXt

CONDOR: COmmunikation in Natürlicher Sprache mit einem (natürlichsprachigen) Dialog-Orientierten RetrievalSystem

CTX: Computergestütztes TeXterschließungssystem

DATEV: DATEnVerarbeitungsorganisation der steuerberatenden Berufe in der Bundesrepublik Deutschland, eingetragene Genossenschaft (Paumgartnerstr. 6-14, 8500 Nürnberg 1)

DB-System: DatenBankSystem

DEUSEM: DEUtsches SEMantisches Wörterbuch im Rahmen des Saarbrücker Analyseverfahrens

DIN: Deutsche Industrie-Norm

DIRS-GRIPS: DIMDI Information Retrieval System-General Relation-Based Information Processing System (entwickelt am Deutschen Institut für Medizinische Dokumentation und Information in Köln)

DUMS: Daten-UMsetzungs-System

DUMS-G: Daten-UMsetzungs-System für GOLEM zur Umsetzung der Deskriptorbasisdaten aus der rechnerunabhängigen Schnittstelle in das erforderliche GOLEM-Eingabeformat

DUMS-T: Daten-UMsetzungs-System für TELDOK (analog zu DUMS-G)

EG: Europäische Gemeinschaft

EIG: EIGenname

EUROTRA: EUROpean TRAnslation System: Bei der Kommission der Europäischen Gemeinschaften in der Entwicklung befindliches europäisches computergestütztes Übersetzungssystem

FALEX: FAchspezifisches LEXikon

FEW: Bezeichnung für "Feste Wendung" in einer Zerlegungsrelation

G: Relator zur Kennzeichnung einer Beziehung zwischen einem Nomen und einem zugehörigen Genitiv-Attribut in einem Komplexen Deskriptor (vgl. GEN-Relation)

GAN: Partitive Relation GANzes, der Ausgangsbegriff enthält das durch das Relatum Bezeichnete

GEG: Relation GEGenbegriff oder Komplement (Antonymrelation)

GEN-Relation: Zur Kennzeichnung einer Beziehung zwischen Nomen und GENitiv-Attribut in einem Komplexen Deskriptor

GOLEM: GroßspeicherOrientierte, Listenorganisierte ErmittlungsMethode.

HYPER: HYPERnym

HYPO: HYPOnym

ID-Relation: In der GOLEM-Feinrecherche verwendete Relation "InformationsDokument" zur Verbindung eines mehrdeutigen Lemmas mit einem besonderen Dokument

IR-System: Information-Retrieval-System (=IRS)

IUS: Relation "Regelungsgegenstand-Juristische Regelung"

JUDO: Forschungsvorhaben JUristische DOkumentanalyse: "Modellentwicklung eines Verfahrens zur computergestützten Indexierung am Beispiel der Analyse juristischer Dokumente"

JUDO-DS: Folgeprojekt zu JUDO zur Evaluierung und anwendungsorientierten Weiterentwicklung des JUDO-Systems für das Fachgebiet DatenSchutzrecht

JUDO-G: Realisierung einer GOLEM-Retrievalkomponente im Projekt

JUDO-T: Realisierung einer TELDOK-Retrievalkomponente im Projekt

JURIS: JURistisches InformationsSystem zur Sammlung, Aufbereitung und Verbreitung von Rechtsinformationen. Eine Entwicklung des BMJ, Bonn

K: Relator zur Kennzeichnung einer Beziehung zwischen einem Nomen und seiner nominalen Anreihung in einem Komplexen Deskriptor (vgl. KON-Relation)

KON-Relation: Zur Kennzeichnung einer Beziehung zwischen Nomen und nominaler Anreihung in einem Komplexen Deskriptor

KOM: Bezeichnung für "Kompositum" in einer Zerlegungsrelation (Ausgangsbegriff)

KWIC-Index: Key-Word-In-Context-Index

LNF: Relation "Langform" zu einem abgekürzten Begriff

MEW: Bezeichnung für "Mehrwortiger Begriff" in einer Zerlegungsrelation

MÜ: Maschinelle Übersetzung

NATURA: Programmpaket zur Analyse einer Problembeschreibung in natürlicher Sprache

NEB: NEBenbegriffsrelation (Kohyponymrelation)

OBR: OBeRbegriffsrelation

P: Relator zur Kennzeichnung einer Beziehung zwischen Nomen und Präpositionalattribut in einem Komplexen Deskriptor (vgl. PRP-Relation )

PARAT: Programmpaket von JURIS zur Überprüfung der Vollständigkeit, der formalen und inhaltlichen Richtigkeit des dokumentarischen Überbaus, zur eigenständigen Ausfüllung bestimmter Rubriken, wobei der Rubrikeninhalt aus anderen Dokumentteilen abgeleitet wird und zur Zerteilung zusammengesetzter Deskriptoren in sinnvolle, d.h. für eine denkbare Abfrage in Betracht kommende Teile dieser Deskriptoren

PASSAT: Programm zur Automatischen Selektion von Stichwörtern Aus Texten. (SIEMENS)

PRESTEL: Britisches System, in der BRD als Bildschirmtext (BTX) bekannt

PRP-Attribut: PRäPositional-Attribut

PRP-Relation: Zur Kennzeichnung einer Beziehung zwischen Nomen und PRäPositional-Attribut

QUA: Quasisynonym-Relation zwischen in der Bedeutung ähnlichen Begriffen

REG: Relation "Juristische Regelung-Regelungsgegenstand"

SADAW: Das SAarbrücker Deutsche Analyse-Wörterbuch. Es enthält die zur Identifizierung von Wortformen nötigen morphologisch-syntaktischen Informationen

SALEM: Saarbrücker Automatische LEMmatisierung

SATAN: Saarbrücker Automatische Text-Analyse. Deutsche Analysekomponente innerhalb des MÜ-Systems SUSY. Entwickelt am SFB 100 Saarbrücken

SFB 100: Sonderforschungsbereich 100 "Elektronische Sprachforschung" an der Universität des Saarlandes, an dem SUSY, das Saarbrücker Übersetzungssystem entwickelt wird

STAIRS: IR-System von IBM

STAIRS-TLS: IR-System von IBM mit Flexionsformenprozessor TLS

SUB: Substantiv

SUSY: Saarbrücker UebersetzungsSYstem

SYN: SYNomymie-Relation

SYS: Synonymierelation zwischen Rechtschreibvarianten

TEI: Relation, Ausgangsbegriff ist "Teil", "Element"

TELDOK: IR-System von Telefunken

TRANSIT: Forschungsvorhaben "TRANSfer von InformationsTechnologie"

TR: Telefunken-Rechner

UBW: Liste der "Unbekannten" Wörter

UNT: UNTerbegriffsrelation

V: Verb

VRB-AKK-Relation: VeRB plus AKKusativkomplement

VRB-MOD-Relation: VeRB plus zugehöriges MODalverb

VWL: VergleichsWortListe: Wörterbuch mit Angaben über die grammatikalische Behandlung und evtl. assoziative Verknüpfung der Wörter bei PASSAT

## Literaturübersicht

Die folgende Zusammenstellung von Literatur zum (weit gefassten) Themenbereich "Automatische Indexierung" schließt ein:

- Literatur, die im Rahmen des Projekts aufgearbeitet (und z.T. in der vorliegenden Untersuchung zitiert) wurde; darunter fallen auch Titel, die speziell nur zum juristischen Bereich gehören;

- Literatur, die im Rahmen von fachbezogenen Lehrveranstaltungen und Seminaren an den Universitäten Regensburg bzw. Saarbrücken behandelt wurde;

- Anfrageergebnisse spezieller Literaturrecherchen, die während des Projektzeitraums durchgeführt wurden (ZDOK, Informationszentrum der GID, LIBRIS, BIBLIODATA, DIALOG).

Agricola, E. (1972): Semantische Relationen im Text und im System. The Hague.

ALPAC-Report. 1966. Languages and maschines, Computers in translation and linguistics. Publication 1416, Automatic Language Processing Advisory Commitee Report. Washington, D.C., Nat. Acad. Sci . Nat. Res. Council.

Alliger, W., Richter, W. (1978) Ein Verfahren zur automatischen Indexierung deutschsprachiger Texte im Rahmen des Internationalen Zweiginformationssystems Elektrotechnik. In: Informatik, Berlin, DDR, 25(1978)4, S. 10-15.

Ammon, R.v. (1976): Erste Ergebnisse einer Analyse zur Deskribierung juristischer Dokumente mit Hilfe von PASSAT. (Arbeitspapier JUDO-A-03), masch., Regensburg.

ANSI Z39.19-1974: American National Standard. Guidelines for Thesaurus Structure, Construction, and Use. American National Standards Institute. New York.

Arbeitsgemeinschaft Information/Dokumentation (1972): Datenverarbeitungsgerechte Indexierung von Informationen. Beiträge zur Fachtagung der Arbeitsgemeinschaft Information/Dokumentation Karl-Marx-Stadt, 13.6.1972. Karl-Marx-Stadt, DDR.

Artandi, S. (1976): Machine Indexing. Linguistic and semiotic implications. In: Journal of the ASIS, Washington, D.C., USA 27(1976)4, S. 235-239.

Arzumanova, I. B., Mamonova, L. A., Pevzner, B. R. (1976): A multilingual automatic indexing System used for simplified machine translation. In: Autom. docum. and math. linguist. New York, USA 10(1976)4, S. 42-44.

AUSBILDUNG (1979): Empfehlung des Sachverständigenkreises 'Ausbildung in Information und Dokumentation' zur Förderung der informationswissenschaftlichen Forschung an Hochschulen. Bonn.

Austin, D. (1976): The role of indexing in subject retrieval. In: HENDERSON 1976, S. 124-156.

Automatische Dokumentbearbeitung im Bereich Rechtsprechung des Juristischen Informationssystems JURIS. Bonn, Mai 1978.

Banerjee, N. (1977): CONDOR: Communication in Natural Language with Dialogue Oriented Retrieval Systems. In: Proceedings of the IFIP Working Conference on Computational Linguistics in Medicine, 2-6 May, 1977, Uppsala.

Banerjee, N. u.a. (1979): CONDOR - Modell eines integrierten Datenbank- und Informationssystems. Siemens AG, München. Bst. Nr. D10/1146-01

Baratta, A. (1971): Gedanken zu einer dialektischen Lehre von der Natur der Sache. In: KAUFMANN 1971, S. 111-118.

Barnes, C. I., Costantini, L., Perschke, S. (1978): Automatic indexing using the SLC-II system. In: In form. process. and managem., Oxford, GB, 14(1978)2, S. 107-119.

Baser, K. H., Cohen, S. M., Dayton, D. L., Watkins, P. B. (1978) On-line indexing experiment at Chemical Abstracts Service. Algorithmic generation of articulated index entries from natural language phrases. In: J. of chem. inform. and comput. sci., Washington, D. C. , USA 18(1978)1 , S. 18-25.

Batori, I.; Krause, S.; Lutz, H.-D. (Hrsg.): Linguistische Datenverarbeitung: Versuch einer Standortbestimmung im Umfeld von Informationslinguistik und Künstlicher Intelligenz, Tübingen: Niemeyer 1982. (Sprache und Information, Bd. 4)

Batten, W. E. (1975): Handbook of Special Librarianship and Information Work. London.

Gatten, W. E. (1977): WURIM 2. Proceedings, Amsterdam, 23.03.-25.03.1976. London, Aslib 1977.

BDSG: Bundesdatenschutzgesetz. Gesetz zum Schutz vor Mißbrauch personenbezogener Daten bei der Datenverarbeitung vom 27.01.1977, BGBL 1, S. 201.

Belzer, J., Stanley, C. E., Prentiss, C. (1974): Regeneration of information rather than information retrieval. Concept creation method. Final report. University of Pittsburgh, Pa., USA.

Bergmann, R. (1977): Homonymie und Polysemie in Semantik und Lexikographie. In: Sprach-

wiss., Heidelberg, S. 27-60.
Bien, J. S., Bolc, L., Szpakowicz, S. (1977): Bootstrapping methodology of computational linguistics research. In: Madey, J. (Hrsg.): Selected topics in information processing IFIP-INFOPOL-76. Proceedings. Warschau, 22.03.-27.03.976. Amsterdam, Niederlande: North-Holland, S. 205-211.
Billmeier, R. (1979): Ein pragmatisch orientiertes linguistisches Analyseverfahren zur automatischen Strukturerkennung für deutsche Sätze: Anwendung eines ATN-Parsers in einem Informationssystem. In: KUHLEN 1979, S. 283-310.
Billmeier, R. (1980): Entwurf einer praxisorientierten Analysegrammatik; automatische Struktur und Inhaltsanalyse für ein nichtrestriktives Information-Retrieval-System. Regensburg 1980.
Bookstein, A., Swanson, D. R. (1974): Probabilistic models for automatic indexing. In: Journal of the ASIS, 25(1974), S. 312-318.
Borko, H. (1977): Toward a theory of indexing. In: Inform. process. and managem., Oxford, Großbritannien, 13(1977)6, S. 235-365.
Braun, S. (1973): Automatische Indexierung durch linguistische Syntaxanalyse. GI-Jahrestagung, Okt. 1973, Berlin, Heidelberg, New York, S. 414-420.
Braun, S. (1974) : Algorithmische Linguistik. Stuttgart.
Braun, S. (1977): Linguistically based methods for indexing and thesaurus construction in information systems. In: Madey, J. (Hrsg.): Selected topics in Information proces sing IFIP-INFOPOL-76. Proceedings. Warschau, Amsterdam, Niederlande: North-Holland, S. 187-203.
Braun, S., Schwind, C. (1975): Automatic, Semantic based Indexing of Natural Language Texts for Information Retrieval Systems. TUM, Report Nr. 7505, 153.
Braun, S., Langendörfer, H. (1976): Aufbau eines Dokumentationssystems für eine Fachbibliothek. In: Hossfeld, F. (Hrsg.): Praxis der Realisierung von Informationssystemen. 4. Workshop und Treffen des German Chapter of the ACM, Jülich, München, S. 119-132.
Brisner, 0. (1975): A system for automatic indexing and keyword translation as the basis for international co-operation. In: Husbands, C. W. (Hrsg.) u.a. (1975): Information revolution. Washington, D.C., USA, S. 75-76.
Broxis, P. F. (1976) : Syntactic and semantic relationships. Or: a review of PRECIS: a manual of concept analysis and subject indexing. In: Indexer, London, 10(1976)2, S. 54-59.
Bruderer, H. E. (1978): Handbuch der maschinellen und maschinenunterstützten Sprachübersetzung. München.
Brühl, G., Meder, W.: Benutzungsseitige Anforderungen an Information-Retrieval-Systeme. Funktionenkatalog. (ZMD-A-30), Berlin.
Brun, W.; Jahn, M.; Stiegler, M. (1979) : Beschreibung und Erläuterung der Identifikations- und Zerlegungsdateien. Bericht JUDO-B-09, masch., Regensburg.
Bundesministerium der Justiz (Hrsg.), (1972): Das Juristische Informationssystem. Analyse, Planung, Vorschläge. Karlsruhe.
Burhenne, W., Perband, K. (Hrsg.) (1970): EDV-Recht. Systematische Sammlung der Rechtsvorschriften, organisatorischen Grundlagen und Entscheidungen der elektronischen Datenverarbeitung.Berlin.
Cercone, N. (1977): Morpho1ogica1 analysis and 1exicon design for natural language processing. In: Comput. and the humanities, 11(1977)4, S. 235-258.
CIRCE (1978): Kommission der Europäischen Gemeinschaften. September 1978. CIRCE, Informations- und Dokumentationssystem der Europäischen Gemeinschaften. Broschüre Nr. IX-1454/78. Brüssel.

Cleverdon, C. W. (1964): Identification of Criteria for Evaluation of Operational Information Retrieval Systems. Cranfield College of Aeronautics, England (zitiert nach SALTON 1968, S. 348).
Cleverdon, C. W. (1977): A survey of the development of information retrieval systems. In: Batten, W. E. (Hrsg.): EURIM 2. Proceedings, Amsterdam, London, S. 103-106.
Clough, C. R. (1978): ASSASSIN. The quiet revolution. In: Program, London 12(1978)1, S. 35-41.
Colby, C. M., Schank, R. (Hrsg.), (1973): Computer Models of Thought and Language. San Francisco.
CONDOR (1978): Sammelband. CONDOR-Veröffentlichungen 1973-1978. München.
Conradi, J. (1973): Steuerrechtsdokumentation. Erfahrungen bei der DATEV. Vortrag beim IBM-Forum 1973 für Wissenschaft und Verwaltung an der Universität Mannheim.
Council of Europe, Documentation Center for Education in Europe (1974): EUDISED Technica1 Studies 1973-1974. Strasbourg, Frankreich.
Craven, T. C. (1976): An experiment in teaching NEPHIS, a nested-phrase indexing System. In: Association Canadienne des Sciences de l' Information: Proceedings of the 4. Canadian conference an information science, London, 11.05.-14.05.1976, Ottawa, Kanada, S. 131-139
Craven, T. C. (1978): Linked phrase indexing. In: Inform. process. and managem., Oxford, Großbritannien, 14(1978)6, S. 469-476.
CTX - Computergestütztes Texterschließungssystem (1982). Kurzbeschreibung des Systems. Veröffentlichungen der Fachrichtung Informationswissenschaft. Saarbrücken.
Dacken, G. (1972): Die Textvorbereitung zur automatischen syntaktischen Analyse russischer Texte. In: Universität Saarbrücken: Zur Strategie einer maschinellen russischen Syntaxanalyse, LA 11. Saarbrücken, S. 25-42.
Dammann, U. (1971): Juristische Dokumentationssysteme und Rechtsentwicklung. In: ZRP 12(1971), S. 287ff.
Davis, C. C. (1978): Reference retrieval by user-negotiated term frequency ordering within a dynamically adjusted national 'document'. In: Journal of informatics, London, Großbritannien, 2(1978)1, S. 62-77.
Debons, A. (Hrsg.), (1974): Information Science. Search for Identity. New York.
Degens, P.O. (1983): Hierarchische Clusteranalyse: Eigenschaften und Berechenbarkeit. In diesem Band.
Dietrich, R. (1972): Wortbildung und automatische Lemmatisierung. In: Universität Saarbrücken: Aspekte der automatischen Lemmatisierung, LA 12, S. 45-61.
Dietze, J. (1974): Ein frequenzstatistisches Verfahren zum automatischen Indexieren. Vortrag auf dem internationalen Symposium 'Das einheitliche System von IRSP innerhalb des ISWTI', Moskau, Sowjetunion, 10.09.-12.09.1974. Berlin, DDR.
Dietze, J. (1977): Ein frequenzstatistisches Verfahren zum automatischen Indexieren. In: Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung, Berlin, DDR, 30(1977)1, S. 58-69.
Dietze, J. (1977): Unterschiede und Gemeinsamkeiten zwischen Thesaurus und Bezugssystem für das automatische Indexieren. In: Wissenschaftliche Zeitschrift der Martin-Luther-Universität Halle-Wittenberg, Halle, DDR, 26(1977)5, S. 105-108.
DIN 1463 (1976): Richtlinien für die Erstellung und Weiterentwicklung von Thesauri. Berlin.
DIN 31623 (1978): Indexierung zur inhaltlichen Erschließung von Dokumenten (Entwurf). Deutsches Institut für Normung e.V. (DIN). Berlin.
DUDEN (1973): Grammatik der deutschen Gegenwartssprache. Band 4. Mannheim.

Dunham, C. S., Pacak, M. G., Pratt, A. W. (1978): Automatic indexing of pathology data. In: Journal of the ASIS, Washington, D.C., USA, 29(1978)2, S. 81-90.
Eberle, C.-E.; Ga rstka, H.; Wegscheider, H. (1976) : Automation juristischer Entscheidungsinstanzen. In: STEINMÜLLER 1976, S. 106-130.
Eggers, H., u.a. (1969): Elektronische Syntaxanalyse der deutschen Gegenwartssprache. Tübingen.
Eggers, H. (1976): Probleme der Identifikationsgrammatik und ihre Anwendung. In: Universität Saarbrücken: Automatische Lexikographie, Analyse und Übersetzung. Saarbrücken, S. 24-29.
Eisenberg, P. (Hrsg.) (1976) : Maschinelle Sprachanalyse. Berlin.
Engel, U. (1977): Syntax der deutschen Gegenwartssprache. Berlin.
Ermer, L. (1974): LEDOC. Ein automatisiertes Dokumentationssystem für Literatur und Rechtsprechung. Beschreibung Siemens System 4004 (1974; Anlage 3).
Evans, L. A., Lynch, M. F., Willett, P. (1978): Structural search codes for on-line compound registration. In: J. of chem. iaform. and comput. sci., Washington, D.C., USA, 18(1978)3, S. 146-149.
Fiedler, H. (1972/1973). Wandlungen der "Automationsgerechten Rechtsgebung". In: DVR Bd. 1(1972/1973), S. 41ff.
Fiedler, H. (1973): Perspektiven juristischer Informationssysteme. In: ÖVD 1973, S. 443ff.
Fiedler, H., Gebhardt, F., Müller, B. S., Poetsch, J. Reiner, G. Stellmacher, 1. (1975): Methodische Erfordernisse juristischer Informationssysteme. Bemerkungen zur Entwicklung von JURIS. In: GEBHARDT 1975.
Field, B. J. (1975): Towards automatic indexing. Relationship between free- and controlled-language indexing and the automatic generation of controlled subject headings and classifications. London.
Field, B. J. (1977): Automatische Indexierung für mehrsprachige Systeme. In: Kommission der Europäischen Gemeinschaften: 3. europäischer Kongress über Dokumentationssysteme und -netze, Luxemburg 3.5.-6.5.1977. München, S. 421-446.
Filipec, J. (1975): Relevante Terminusmerkmale und einige Möglichkeiten ihrer Bearbeitung durch technische Mittel. In: Wiss. Z. D. Tech. Univ. Dresden, 24(1975)6, S. 1249-1252
Fisher, J. N. 0. (1977): Information and the explicitly perfomative verb. In: J. of Informatics, London, 1(1977)1, S. 3-16.
Frankenberger, R. (1978): Bibliothekarische Sacherschließung. Neue Aspekte für die Benutzer durch Einsatz der EDV und Rückwirkungen auf den Bibliotheksbetrieb. In: HABERMANN 1977, S. 202-209.
Frei, H. P. (1977): Automatische Klassifikationsmethoden. In: Nachr. nouv. not., Bern, Schweiz, 53(1977)6, S. 308-314.
Fugmann, R. (1979): Die Aufgabenteilung zwischen Wortschatz und Grammatik in einer Indexsprache. In: KUHLEN 1979, S. 67-94.
Fuhr, N. (1983): Klassifikationsverfahren bei der Automatischen Indexierung. In diesem Band.
Gebhardt, F. (Hrsg), (1975): Beiträge zur Methodik juristischer Informationssysteme. Berlin.
Gebhardt, P. (1977): Wortstatistik an größeren Textsammlungen. In: Nachr. f. Dokum., 28(1977)2, S. 53-57.
Gerhardt, T. (1978): Kodieranweisung für die SESAM-Wörterbücher. Sonderforschungsbereich Elektronische Sprachforschung, masch., Saarbrücken.
Glück, R. S. (1976): Wirtschaftlichkeit juristischer Informationssysteme. Nutzen-Kosten-Untersuchung zum geplanten Juristische.n Informationssystem für die BRD. Beiträge zur juristischen Informatik 6(1976). Darmstadt.

Gnany, R. (1976): A computer-assisted indexing procedure based on the AGRIS input format. In: Quart. bull. of the Intern. Assoc. of Agric. Librarians and Documentalists, Bennekom, Niederlande, 20(1976)3/4, S. 128-135.
Graichen, D. (1981): Thesaurusunabhängiges Indexieren medizinischer Befunde mit "INDEX2". In: Informatik 28 (1981) 4, S.30-35.
Grishman, R. (1973): Implementation of the string parser of English. In: Rustin, R. (Hrsg.): Natural language processing. New York, S. 89-109.
Haag, K. (1971): Kritische Bemerkungen zur Normlogik. In: KAUFMANN 1971, S. 135-146.
Haake, R. u.a. (1979): Handbuch der Information und Dokumentation. Leipzig.
Habel, 8. (1979): IBS als Beispiel für den Einsatz verschiedener Informations- und Datenbanksysteme. In: KUHLEN 1979, S. 195-219.
Habel, C., Schmidt, A., Schweppe, H. (1977): On automatic paraphrasing of natural language expressions. Berlin.
Habel, C.; Rollinger, C.-R. (1979): Eine juristische Anwendung zur Repräsentation des BDSG. In: SCHNEIDER 1979, S. 178-194.
Habermann, A. (Hrsg.) u.a. (1977): Die wissenschaftliche Bibliothek 1977. 67. Deutscher Bibliothekstag, Bremen, 31.05.-04.06.1977. Frankfurt.
Haft, F. (1972): Automatisierte juristische Dokumentation und Gesetzesplanung. In: EDV und Recht, Bd. 4 (1972), Berlin.
Hahn, W. v. (1978): Überlegungen zum kommunikativen Status und der Testbarkeit von natürlichsprachlichen Artificial-Intelligence-Systemen. In: Sprache und Datenverarbeitung 1(1978), S. 3-16.
Haller, J. (1978): Ein Algorithmus zur Reduktion syntaktischer Homographien in einem Informationssystem. Diss., Regensburg.
Hammer, D. P. (Hrsg.) (1976): The information age, its development, its impact. Metuchen, N.J., USA.
Hann, M. L. (1977): Computers and the Production of Systematic Terminological Glossaries. In: ALLC Bulletin 5(1977), S. 26 ff.
Harris, B. (1976): Faceting. Paper presented at Varna, Bulgarien, 05.1975. In: TA inform., Paris, Frankreich, 17(1976)2, 44-50.
Harter, S. P. (1978): Statistical approaches to automatic indexing. In: Drexel libr. quart., Philadelphia, USA, 14(1978)2, S. 57-74.
Hartmann, F. (1976): Das Text Retrieval System/1. Aufgaben, Aufbau und Arbeitsweise. In: GMD-Spiegel, St. Augustin, 21976), S.9-18.
Heaps, H. S. (1978): Information Retrieval. Computational and Theoretical Aspects. New York.
Heer, T. (1979): Quasi comprehension of natural language simulated by means of information traces. In: Inform. process. and managem., Oxford, 15(1979)2, S. 89-98.
Heger, K. (1971): Monem, Wort und Satz. Tübingen.
Heger, K.; Petöfi, J. S. (Hrsg) (1977): Kasustheorie, Klassifikation, semantische Interpretation. Hamburg.
Heinemann, R. (1971): Überblick über einige wesentliche Methoden und Systeme zur automatischen Indexierung von vollständigen Texten. Abschlussarbeit, Technische Hochschule Ilmenau, Institut für Informationswissenschaft, Erfindungswesen und Recht (INER), Ilmenau, DDR.
Helbich, J. (1978): Some results of an experiment in statistical selection of keywords. In: HORECKY 1978, S. 159-175.
Hellwig, P. (1977): Dependenzanalyse und Bedeutungspostulate. Heidelberg.
Herderson, K. L. (1976): Major Classification Systems. Urbana, I11., USA.

Hendrix, G. G., Sacerdoti, E. D., Sagalowicz, D., Slocum, J. (1978): Developing a natural language Interface to complex data. In: ACM trans. on database syst., New York, 3(1978)2, S. 105-147.
Henke, H. (1965): Untersuchungen über das maschinelle Erkennen flektierter Wortformen. Arbeitsblätter des Lehrstuhls für Elektronische Rechenanlagen der TH Hannover, Nr. 13.
Hirschmann, L. , Grishman, R. , Sager, N. (1976) : From text to structured information - automatic processing of medical reports. In: Winkler, S. (Hrsg.) u.a.: 1976 national Computer conference. New York 07.06.-10.06.1976, Amer. Feder. of Inform. Process. Soc. (AFIP.S), Montvale, N.J., USA, AFIPS press, S. 267-275.
Hitzenberger, L. u.a. (1977): Das Regensburger Textverarbeitungssystem COBAPH. In: Sprache und Datenverarbeitung 1(1977), 5.30-32.
Hofmann, J. (1977): Satzexterne freie nicht-referentielle Verweisformen in juristischen Normtexten. (Zulassungsarbeit), masch., Regensburg.
Hoffmann, D. u.a. (1971): Automatische Textanalyse mit PASSAT. Überlegungen, Versuche und erste Ergebnisse. In: Zeitschrift für Datenverarbeitung, S. 495-504.
Horecky, H. (Hrsg.) u. a. ( 1978) : Prague Studies in Mathematical Linguistics. Vol. 6. Amsterdam.
Horn, D. (1971): Die semantischen Aspekte der Informationswiedergewinnung. In: NJW 1971, S. 1588 ff.
Horn, D. (1974): Computereinsatz im Rechtswesen. In: DSWR 1974, S. 56 ff.
Hoppe, A. (1976): Dokumentation und maschinelle Text-InhaltAnalyse. In: HARBECK, S. 83-106.
Husbands, G. W. (Hrsg.) u.a. (1975): Information Revolution. 38. ASIS Annual Meeting, Boston, Mass., USA, 26.10.-30.10.1975, Washington D. C.
Hutchins, W. I. (1970): Linguistic Processes in the Indexing and Retrieval of Documents. In: Computer Physics Communication, S. 29-64.
Hutchins, W. I. (1975): Languages of Indexing and Classification. London.
Hutchins, W. I. (1976): Languages of Indexing and Classification. A Linguistic Study of Structure and Functions. Stevenage.
INFORMATIONSLINGUISTIK (1979): Empfehlung des ad-hoc-Ausschusses des Bundesministeriums für Forschung und Technologie zur Informationslinguistik. Bonn.
Informationssysteme für die 80er Jahre. Referate der 2. Gemeinsamen Fachtagung der Östereichischen Gesellschaft für Informatik (ÖGI) und der Gesellschaft für Informatik (GI), Bd. 1,2. Linz/Österreich 1980.
Ivankin, V. I. (1976): Algorithmic evaluation of keyword selection techniques in coordinate indexing. In: Autom. docum. and math. linguist., New York, N.Y., USA, 10(1976)2, S. 43-50.
Jahoda, G. , Foos, F. A. ( 1971) : The use of an online searched and printed coordinate index in teaching. Tallahassee, Florida, USA.
Jaene, H., Seelbach, D. (1975): Maschinelle Extraktion von zusammengesetzten Ausdrücken aus englischen Fachtexten. Berlin, Köln, Frankfurt.
Jannucci, B. I., Ragona, B., Taddei-Elmi, G. (1978): Notes sur un Système de question et de recherche automatique d'informations juridiques. In: Institute de Recherche d' Informatique et d'Automatique (1978): International seminar on intelligent question-answering and data base systems. Proceedings, Bonas, Frankreich, 21.6.-30.6.1977, Le Chesnay, Frankreich, S. 311-320.
Janos, J. (1976): Vysledky experimentalniho overeni metod automatizovane komprimace odbornych textu v anglictine, nemcine, rustine, a francouzstine (Ergebnisse der Versuchsüber-

prüfung von Methoden der automatisierten Verdichtung von Fachtexten in Englisch, Deutsch, Russisch und Französisch). In: Cescoslovenska Inform., Prag, Tschechoslowakei, 18(1976)12, S. 330-333.
Janos, J. (1978): Results of an experiment in automatic extracting and problems in the use of condensed texts in automated information systems. In: Int. forum an inform. and docum., Den Haag, Niederland, 3(1978)1, S. 13-17.
Jones, A. (Hrsg.) (1976): The Computer in 1iterary and 1inguistic studies. Cardiff, GB, Univ. of Wales Press.
Jordan, S. R. (1972): Learning to use contextual patterns in language processing. Madison, Wis., USA.
Joshi, A. K., Weischedel, R (1977): Computation of a subclass of inferences: presupposition and entailment. In: Amer. J. of comput. linguist. microfiche, Menlo Park, Calif., USA.
JUDO. Juristische Dokumentanalyse. Forschungsbericht (1977 - 1979). Regensburg 1980.
JURIS (1972): Bundesministerium der Justiz (Hrsg.): Das Juristische Informationssystem - Analyse, Planung, Vorschläge. Karlsruhe.
JURIS (1978): Juristisches Informationssystem. Projektbeschreibung. Bonn.
JURIS-Benutzertagung (1979): Der Bundesminister der Justiz: Ergebnisniederschrift des 4. Erfahrungsaustausches der JURIS-Benutzer am 25./26.09.1979. Bonn.
Kaplan, R. M. (1973): A general syntactic processor. In: RUSTIN 1973, S. 193-241.
Kassel, H.; Strnad, P. (1978): Lexikon Datenschutz und Datensicherung. Berlin, München.
Kaufmann, A. (Hrsg.) (1971): Rechtstheorie, Ansätze zu einem kritischen Rechtsverhältnis. Karlsruhe.
Kaufmann, A.; Hassemer, W. (1971): Grundprobleme der zeitgenössischen Rechtsphilosophie und Rechtstheorie. Frankfurt/Main.
Kay, M. (1970): The MIND System. In: Rustin, R. (Hrsg.): Natural language processing. New York, S. 155-188.
Keen, E. M. (1971): An analysis of the documentation requests. In: SALTON 1971, S. 181-205.
Keller, H. (1973): Establishing a German Root System by Computer. In: Computers and the Humanities, 7(1973), S. 199ff.
Kelly, E. F., Stone, P.J. (1975): Computer recognition of English word senses. Amsterdam.
Kelly, I. D. K. (1977): PROTRAN - ein Instrument mit allgemeinem Anwendungsbereich für die Übersetzung natürlicher und algorithmischer Sprachen. In: Kommission der Europäischen Gemeinschaften: 3. europäischer Kongress über Dokumentationssysteme und -netze, Luxemburg, 03.05.-06.05.1977. München, (1977) Bd. 1, S. 579-593.
Kelsen, H. (1960): Reine Rechtslehre. 2. vollst. neu bearb. und erw. Aufl. (l. Auflage: 1934) Wien.
Kirsner, Z., Buraneva, E. (1976): A method of automatic indexing. In: Autom. docum. and math. linguist., New York, N.Y., USA, 10(1976)3, S. 92-95.
Kittredge, R. (1976): Transformational decomposition and transfer grammar. In: TA inform., Paris, 17(1976)2, S. 50-54.
Klein, W., Rath, R. (1971): Automatische Lemmatisierung. Saarbrücken.
Knaack, J. (1979): Eine Pressedatenbank für Text. Erschließung und Suche von heterogenen Inhalten mit differenziertem Vokabular. In: KUHLEN 1979, S. 171-194.
Koller, G., (1975): Syntaktische Analyse von Texten natürlicher Sprachen im Dialog Mensch Maschine. Hamburg.
Kopelent, J. (1978): Zur Auflösung semantischer Mehrdeutigkeiten bei Verben im Informationsrecht. (Arbeitspapier JUDO-A-12), masch., Regensburg.
Kopelent, J. (1979): Erste statistische Auswertung der Wahrscheinlichkeitsgewichtungen bei den

Bedeutungsvarianten mehrdeutiger Substantive, Verben und Adjektive des Informationsrechts. (JUDO-V-03), masch., Regensburg.
Kopelent, W. (1978): Zum Zusammenhang von Nominalkomposita und entsprechenden verbhaltigen Strukturen am Beispiel juristischer Fachtexte. (Magisterarbeit), masch., Regensburg.
Korolev, A. E. u.a. (1977): The statistical and lexicogrammatical properties of words. In: Autom. docum. and math. 1inguistics, New York, 11(1977)1, S. 1-11.
Korolev, E. 1. (1977): The use of the distributive statistical method in the language apparatus of automated information Systems. In: Autom. docum. and math. linguist., New York, 11(1977)1, S. 31-37.
Krallmann, D. (1966): Statistische Methoden in der linguistischen Textanalyse. Diss. , IKP, Bonn.
Krallmann, D. (Hrsg.) (1978): Kolloquium zur Lage der linguistischen Datenverarbeitung. LDV-Fittings, 22.2.-24.2.1978, Essen.
Krause, J. u. a. (1976) : Programmsystem COBAPH. 2. Aufl., masch., Regensburg.
Krause, J., Hitzenberger, L., Konrad, W., Scharinger, K., Schneider, C. (1976): Programmsystem COBAPH. masch., Regensburg.
Kreutz, E., Menetre, E. (1978): CONDOR, a natural-language-based Information system. In: Institut de Recherche d'Informatique et d'Automatique (1978): International seminar an intelligent question-answering and data base systems. Proceedings, Bonas, Frankreich, 21.6.-30.6.1977. Le Chesnay, Frankreich, (1978), S. 199-203.
Kroupa, E. (1979): Kodieranweisungen zur SADAW-Kodierung nach Flexionsmustern. (Arbeitspapier JUDO-B-10), masch., Regensburg.
Kroupa, E. (1982): Strategien zur Dokumentrepräsentation bei CTX. Ein Verfahren zur computergestützten Texterschließung und Textwiedergewinnung. In: Batori, I.; Krause, S.; Lutz, H.-D. (Hrsg.) 1982, S. 155-161.
Krylov, Y. K. u.a. (1977): Statistical Analysis of Polysemy as a Language Universal and the Problem of the Semantic Identity of the Word. In: Autom. Docum. and Math. Linguistics. New York, 11(1977)1, S. 80-87.
Kuhlen, R. (1976): Anforderungen der Informationswissenschaft an die Linguistik . Vortrag gehalten auf der Tagung ' Informatik und Informationswissenschaft', St. Augustin, 13.4.-14.4.1976.
Kuhlen, R. (1977a): Experimentelle Morphologie in der Informationswissenschaft. München.
Kuhlen, R. (1977b): Funktionale Grammatiken und Weltwissen. In: Sprache und Datenverarbeitung, 2(1977), S. 165-171.
Kuhlen, R. (Hrsg.) ( 1979) : Datenbasen, Datenbanken, Netzwerke. München. (bisher erschienen Band 1 und 2)
Kuno, S., Oettinger, A. G. (1963): Multiple-Path Syntactic Analyser. Math. Ling. Aut. Transl. Rpt. NSF8, Harvard, Cambridge, Mass., USA.
Kuno, S. (1976): Automatische Analyse natürlicher Sprachen. In: EISENBERG 1976, S. 167-203.
Kurtz, P., Lowe, T. C., Bakert, T. A. (1971): On-line retrieval 2. Bethesda, Md. , USA.
Laliberte, M. (1977): Quelques problemes rencontres dans l'application de precis a la langue Francaise. In: Can. j. inform. sci., Ottawa, Kanada, 2(1977)1, S. 79-92.
Lambert, G. (1976): PRECIS in a multilingual context - part 4. In: Libri, Kopenhagen, 26(1976)4, S. 302-323.
Lancaster, F. W. (1976): The relevance of linguistics to information.science. In: FID, Comittee on linguistics in documentation (1976): Workshop in linguistics and information science,

Stockholm, Schweden, 3.5.-5.5.1976.
Lancaster, F. W., Owen, J. M. (1976): Information retrieval by computer. In: HAMMER 1976, S. 1-33.
Lancaster, F. W. (1977): Vocabu1ary contro1 in information retrieval systems. In: Advances in librianship. New York, 7(1977), S.2-40.
Lansky (1973): Entschließung zur Erstellung eines einheitlichen Thesaurus und einer einheitlichen juristischen Systematik für die Bundesrepublik Deutschland. In: Nachr. Dok. 24(1973), S.127ff.
Landry, B. C., Rush, J. E. (1975): Automatic indexing. Progress and prospects. In: Belzer, J. (Hrsg.) u. a. (1975) : Encyclopedia of computer science and technology, New York, N.Y., USA, Vol. 2, S. 403-447.
Lehmann, H. (1978): Interpretation of natural in an information System. In: IBM j. of res. and dev., Armonk, N.Y., USA, 22(1978)5, S. 560-572.
Lenders, W. (1975): Semantische und argumentative Textdeskription. Hamburg.
Leont'eva, N. N., Vishnyakova, S. M. (1977): An experiment in automatic indexing with semantic compression. In: Autom. docum. and math. linguist., New York, N.Y., USA, 11(1977)3, S. 26-35.
Lin, C. H. (1978a): SEFLIN - Separate Feature Linear Notation System for chemical compounds. In: J. of chem. inform. and comput. sci. , Washington, D. C. , USA, 18(1978)1, S. 41-47.
Lin, C. H. (1978b): Chemica1 inference based an SEFLIN. 1. Basic cognizance of molecular shape, fragments, and atomic environment of organic compounds. In: J. of chem. inform. and comp. sci. , Washington, D. C. , USA, 18(1978)1 , S. 47-51.
Liston, D. M., Howder, M. L. (1977): Subject analysis. In: Annu. rev. of inform. sci. and technol., Washington, D.C., USA, 12(1977), S. 81-118.
Lustig, G. (1969): Die automatische Zuteilung von Schlagwörtern des EURATOM-Thesaurus. In: Neue Technik, 11(1969), S. 247-256.
Lustig, G. (1972): Probleme der Textverarbeitung bei der automatischen Indexierung. In: SCHANZE 1972.
Lustig, G. (1974): Probleme der Wörterbuchentwicklung für das automatische Indexing und Retrieval. In: Nachrichten für Dokumentation 25, S. 50-54.
Lustig, G. (1979): Ansätze einer realistischen automatischen Indexierung unter Verwendung statistischer Verfahren. In: KUHLEN 1979, S. 339-368.
Lutz, Th., Klimesch, H. (1971): Die Datenbank im Informationssystem. München, Wien.
Lynch, M. F., Willett, P. (1978): The production of machine-readable descriptions of chemical reactions using Wiswesser Line Notations. In: J. of chem. inform. and comput. sci., Washington, D.C., USA, 18(1978)3, S. 149-154.
Lyons, J. (1972): Einführung in die moderne Linguistik. München. (übers. von Introduction to Theoretical Linguistics, Combridge 1968).
Maas, H. D. (1977): Ergebnisse der Satzanalyse und transformationelle Synthese im Saarbrücker Übersetzungssystem SUSY. Sonderforschungsbereich Elektronische Sprachforschung, LA 24, Saarbrücken.
Maas, H. D. (1978): Das Saarbrücker Übersetzungssystem SUSY. In: Sprache und Datenverarbeitung, 1(1978), S. 43-61.
MacAllister, C. (1971): A study and model of machine-like indexing behaviour by human indexers. Los Gatos, Californien, USA.
Mahapatra, M. (1978): Syntactical difference between POPSI und PRECIS. In: Libri, Kopenhagen, Dänemark, 28(1978)3, S. 235-245.
Marcyszewski, W. (1976) : From the concept of the topic of a sentence to the concept of key-

word. Remarks an a research program. In: Autom. docum. and math. Linguist., New York, N. Y., USA, 10(1976)4, S. 64-74.
Maron, M. E., Kuhns, J. L. (1960): On relevance, probabi1istic indexing and information retrieval. In: Journal of the ACM, 7(1960), S. 216-244.
Martin, J. A. R. San (1978): Sistema KWIT (Key-Words in Title) des permutacion selectiva con prenotacion de carga informativa para indizacion de documentacion cientificotecnica. System KWIT (Key-Words in Title) zum Indexieren wissenschaftlich-technischer Dokumentation. In: Rev. esp. de docum. cient. , Madrid, 1(1978)2, S. 149-157.
Messmer, G. (1971): Möglichkeiten der Automatisierung der Rechtsdokumentation. In: Nachr. Dok. 22(1971), S. 191ff.
Meulen, W. A. v. d., Janssen, P. J. F. C. (1977) : Automatic versus manual indexing. In: Inform. process. and managem., Oxford, Großbritannien, 13(1977)1, S. 13-21.
Meunier, J. G. (1976): The Lemmatization of contemporary French. In: Jones, A. (Hrsg.) u. a.: The Computer in literary and linguistic studies. Proceedings. 3. International symposium. Cardiff, GB: Un iv. of Wales Press, S. 208-214.
Meunier, J.G., Rolland, S., Daoust, F. (1976): A system for text and content analysis. In: Comput. and the humanities 10(1976)5, S. 281-286.
Michiels, A., Mullenders, J., Noel, J. (1977): Automatic skimming. The 'LOUISA' System. In: ALLC bull., Stockport, Großbritannien, 5(1977)1, S. 2-14.
Moscovoj, V. A. (1977): Zur Aufhebung der formalen Wortfolge im deutschen Satz. In: Kommission der Europäischen Gemeinschaften: 3. europäischer Kongress über Dokumentationssysteme und -netze, Luxemburg, 03.05.-06.05.1977. Bd. 1. München, S. 561-562.
Moser, A. (1977): Zur Analyse und Bewertung informationeller Prozesse und Systeme. Stuttgart - Bad Cannstadt.
Moyne, J. A. (1975): Relevance of computer science to linguistics and vice versa. In: Int. of comput. and inform. sci., London 4(1975)3, S. 265-279.
Mueller, M. (1961): Zeit-, Kosten- und Wertfaktoren der Informationserschließung. In: IBM-Nachrichten, 149(1961), S. 1348ff.
Müller, B. S. (1976a): Kompositazerlegung. In: MÜLLER 1976b, S. 83-127.
Müller, B. S. (Hrsg.) (1976b): Beiträge zur Sprachverarbeitung in juristischen Dokumentationssystemen. Berlin.
Müller, B. S. (1977): Zum Problem von Frage-Antwort-Systemen im Rechtswesen. In: Sprache und Datenverarbeitung 2(1977), S. 133-140.
Mullenders, J., Noel, J. (1976a): LOUISA (linguistically oriented understanding and indexing System for abstracts). In: Cah. de la Docum., Brüssel, 30(1976)3, S. 81-104.
Mullenders, J., Noel, J. (1976b): LOUISA. Linguistically oriented understanding and indexing system for abstracts. In: FID, Comittee an linguistics in documentation (1976): Workshop in linguistics and information science, Stockholm, Schweden, 3.5.-5.5.1976.
Naumann, P. (1978): Zur Implementierung eines Verfahrens zum automatischen Indexieren von Referaten im Rahmen des Informationsrecherchesystems AIDOS. In: Informatik, Berlin, DDR, 25(1978)2, S. 32-35.
Newell, A., Cooper, F. S., Forgie, J. W., u.a. (1975): Considerations for a follow-on ARPA research program for Speech understanding systems. Pittsburgh, PA., USA.
Niedermayer, W. (1976): Dokumentation und Information Retrieval. In: Kehr (Hrsg.) u.a. (1976): Zur Theorie und Praxis des modernen Bibliothekswesens, München, Bd. 2, S. 268-312.
Noel, J. (1975): Document analysis algorithms and MT research. In: Rev. des Lang. vivantes, Brüssel, 41(1975)3, S. 237-260.
Norwood, J. W. (1975): Machine-aided retrieval. Alexandria, USA.

Oettinger, A. (1960): Automatic Language translation. Cambridge, Mass., USA.
Oettinger, A. (1969): Run, computer, run. Cambridge, Mass., USA.
Olney, J., Lam, V., Yearwood, B. (1976) : A new technique for detecting patterns of term usage in text corpora. In: In form. process. and managem., Oxford, Großbritannien, 12(1976)4, S. 235-250.
Orekhov, Y. V. (1977): Automatic resolution of wordform homonymy. In: Autom. docum. and math. linguist., New York, N.Y., USA, 11(1977)2, S. 35-38.
Panyr, J. (1978): STEINADLER - ein Verfahren zur automatischen Deskribierung und zur automatischen Dokumentklassifikation. In: Nachrichten für Dokumentation, 4/5(1978), S. 184-191.
Pao, M. L. (1978): Automatic text analysis based on transition phenomena of word occurrences. In: Journal of the ASIS, Washington, D.C., USA, 29(1978)3, S. 121-124.
Parkison, R. C., Colby, K. M., Faught, W.S. (1977): Conversational language comprehension using integrated patternmatching and parsing. In: Artif. intell., Amsterdam, 9(1977)2, S. 111-134.
Petöfi, J. S., Rieser, H. (Hrsg.) (1973): Studies in Text Grammar. Dordrecht.
Petöfi, J. S. (1975a): Some problems of text typology and text processing an the basis,of partial text theory. In: PETÖFI 1975, S. 61-91.
Petöfi, J. S. u.a. (1975b): Fachsprache - Umgangssprache. Kronberg.
Petöfi, J. S. (1977): Textrepräsentation und Lexikon als semantische Netzwerke. In: HEGER/ PETÖFI 1977.
Pierce, J. C. (1978): Back-of-book subject indexing with APL. Automated indexing for those without computer background. In: Inform. process. and managem., Oxford, Großbritannien, 14(1978)2, S. 85-91.
Plesch, M., Griese, J. (1972): Eigenschaften von Datenbanksystemen - ein Vergleich. In: Angew. Informatik 11(1972), S.489-498.
Podlech, A. (1972/1973): Verfassungsrechtliche Probleme öffentlicher Informationssysteme. In: DVR 1(1972/1973), S. 149ff.
Preschel, B. M. (1972): Indexer consistency in perception of concepts and in choice of terminology. Final report. New York, N. Y. , USA.
Preschel, B. M. (1977): A US indexer attends a PRECIS indexing workshop. In: Indexer, London, 10(1977)3, S. 111-115.
Prestel, B. M. (1971): Datenverarbeitung im Dienste der juristischen Dokumentation. In: EDV und Recht, Bd. 3 (1971), Berlin.
Reisinger, L. (1972): Automatisierte Normanalyse und Normanwendung. Arbeitspapiere Rechtsinformatik, Heft 7. Berlin.
Reisinger, L. (1973): Die automatisierte Messung juristischer Begriffe. Arbeitspapiere Rechtsinformatik, Heft 9. Berlin.
Richmond, P. A. (1976): Classification from PRECIS. Some possibilities. In: J. of the ASIS, Washington, D.C., USA, 27(1976)4, S. 240-247.
Rinewalt, R. J. (1977): Feature Evaluation of a Full Text Information Retrieval System. In: On-Line Review 1, 43.
Robertson, S. E. (1977): Theories and models in information retrieval. In: J. of docum., London, 33(1977)2, S. 126-148.
Robertson, S. E. (1978): Indexing theory and retrieva1 effectiveness. In: Drexel libr. quart., Philadelphia, USA, 14(1978)2, S. 40-56.
Rostek, L. (1979): Methoden des partiellen Parsing für das automatische Indexing - Syntaxgraphen zur Analyse von Sprachmustern. In: KUHLEN 1979, S. 251-282.

Rothkegel, A. (1972a): Zur semantischen Subkategorisierung. In: Universität Saarbrücken: Aspekte der automatischen Lemmatisierung. LA 12, Saarbrücken, S.11-31.
Rothkegel, A. (1972b): Feste Syntagmen. In: Universität Saarbrücken: Aspekte der automatischen Lemmatisierung. LA 12, Teil II, Saarbrücken, S. 95-137.
Rothkegel, A. (1976): Valenzgrammatik I. Sonderforschungsbereich Elektronische Sprachforschung, Universität Saarbrücken. Linguistische Arbeiten 19. Saarbrücken.
Rustin, R. (Hrsg.) (1973): Natural language processing. New York.
Sager, N. (1973): The string parser for scientific literature. In: Rustin, R. (Hrsg.): Natural language processing, New York, S. 61-87.
SALEM (1980): Ein Verfahren zur Automatischen Lemmatisierung Deutscher Texte. Hrsg.: Sonderforschungsbereich 100 "Elektronische Sprachforschung", Projektbereich A. Tübingen: Niemeyer.
Salton, G. (1968): Automatic Information Organization and Retrieval. McGraw-Hill. New York.
Salton, G. (Hrsg.) (1970): Information storage and retrieval. Reports on analysis, dictionary construction, user feedback, clustering, and on-line retrieval. Ithaca, N.Y., USA.
Salton, G. (Hrsg.) (1971): The SMART Retrieval System. Experiments in Automatic Document Processing. Englewood Cliffs, N.J., USA.
Salton, G. (1972a): Experiments in automatic thesaurus construction for information retrieval. Information Processing 1971, North Holland, Amsterdam.
Salton, G. (1972b): A new comparison between conventional indexing (Medlars) and automatic text processing (SMART). In: Journal of the ASIS, 23(1972), S. 75-84.
Salton, G. (1975): System Testing. In: Salton, G. (1975): Dynamic Information und Library Processing. Englewood Cliffs, N. J., USA.
Salton, G.(1981): A Blueprint for Automatic Indexing. In: SIGIR Forum 16 (1981) 2, S. 22-38.
Salton, G., Lesk, M. E. (1971a): Computerevaluation of indexing and text processing. In: Salton 1971, S. 143-180.
Salton, G.; Lesk, M. E. (1971b): Information Analysis and Dictionary Construction. In: SALTON 1971, S. 113-142.
Salton, G., Wong, A. (1976): On the role of words and phrases in automatic text analysis. In: Comput. and the humanities, 10(1976)2, S. 69-87.
Salu, L. (1977): PRECIS (PREserved Context Indexing System) In: Open, Den Haag, Niederland, 9(1977)1, S. 19-29.
SATAN-Handbuch (1978ff): Sonderforschungsbereich Elektronische Sprachforschung. Teilprojekt A1: Automatische Lemmatisierung. Loseblattsammlung, Saarbrücken.
Schaeder, S. (1978): Maschinelle Dokumentation und Lexikographie. In: Krallmann, D. (Hrsg.): Kolloquium zur Lage der linguistischen Datenverarbeitung. Essen, 22.02.-24.02.1978, LDV-Fittings (1978) S. 110-164.
Schank, R. (1973): The conceptual analysis of natural language. In: RUSTIN 1973, S. 291-309.
Schanze, H. (Hrsg.) (1972): Literatur und Datenverarbeitung. Tübingen.
Schmidt, E. L., Thiel, M. (1972): Zur Behandlung nicht inventarisierter Komposita. In: Universität Saarbrücken: Aspekte der automatischen Lemmatisierung. LA 12, Teil II. Saarbrücken, S. 82-94.
Schmitz-Esser, W. (1977): Die Pressedatenbank für Text und Bild des Verlagshauses Gruner & Jahr. In: Nachrichten für Dokumentation, 3(1977), S. 124-131.
Schneider, H.-J. u.a. (1979): Automatische Erstellung semantischer Netze. TU Berlin, BMFT-FB ID 79/05.
Schneider, W., Sagvall Hein, A. (Hrsg.) (1977): Computational linquistics in medicine. Amsterdam.

Schneider, H.-J. (1979): Automatische Erstellung semantischer Netze. TU Berlin. (BMFT-FB ID 79/05).
Schneider, Ch. (1976): Probleme der Nominalisierung (satzübergreifende Verknüpfungen). (Zulassungsarbeit), masch., Regensburg.
Schneider, Ch. (1979a): Die Anwendung linguistischer Verfahren bei der Analyse juristischer Dokumente. Vortrag: Tagung der SIW, Heidelberg, 7./8.6.79. (JUDO-V-09, masch., Regensburg), (erscheint in: KUNZ (Hrsg.): SIW-Tagungsband).
Schneider, Ch. (1979b): Lemmatisierung im Projekt JUDO. (JUDO-V-08), masch., Regensburg. (erscheint im ALLC-Bulletin).
Schneller, H. (1965): Automatische Übersetzung von Sprachen. Stand und Tendenzen der gegenwärtigen Forschung. In: FRANK 1965, S. 254f f.
Schott, G. (1971/1972): Automatic Analysis of Inflectional Morphems in German Norms. An Algorithm for Automatic Indexing. In: Acta Informatica 1(1971/1972), S. 360ff.
Schott, G. (1978): Automatische Kompositazerlegung mit einem Minimalwörterbuch zur Informationsgewinnung aus beliebigen Fachtexten. In: WINGERT 1978, S. 32-43.
Schrimpf, J. (Hrsg.) (1973): IBM-Beiträge zur Datenverarbeitung. Stuttgart.
Schüter, M. (1971): Methoden zur Indexierung vollständiger Texte; ihre Vor- und Nachteile. Abschlußarbeit, Technische Hochschule Ilmenau, Institut für Informationswissenschaft, Erfindungswesen und Recht (INER), Ilmenau, DDR.
Schulze, U. (1977): Untersuchungen zur automatischen Erzeugung eines Vokabulars aus Texten juristischer Entscheidungen und zur Strukturierung dieses Vokabulars durch Anwendung von Clusterverfahren. Gesellschaft für Mathematik und Datenverarbeitung (GMD), GMD-Mitteilungen 40, Bonn
Schwuchow, W. (1970a): Zur Messung der Wirtschaftlichkeit von Dokumentationsnachweissystemen. München-Pullach, Berlin.
Schwuchow, W. (1970b): Über die Bewertung von Dokumentationssystemen. In: Nachr. Dok. 21(1970), S. 239ff.
Schwuchow, W. (1971): Benutzeranalysen als Grundlage für die Organisation von Informations- und Dokumentationseinrichtungen. In: Nachr. Dok. 22(1971), S. 237ff.
Schwuchow, W. (1972): In welchem Umfang ist die Wirtschaftlichkeit von Dokumentationssystemen meßbar? In: Nachr. Dok. 23(1972), S. 7ff.
Sebiger, H. (1973): Erfahrungsbericht über den Aufbau einer Steuerrechtsdokumentation. Referat beim IBM-Seminar "Informationssysteme in Regierung und Verwaltung", Bad Liebenzell, November 1971. In: SCHRIMPF 1973, S. 69ff.
Seelbach, D. (1975): Computerlinguistik und Dokumentation. Key Phrases in Dokumentationsprozessen. München.
Seelbach, H. E. (1977): Von der Stichwortliste zum halbautomatisch kontrollierten Wortschatz. In: Nachr. f. Dokum., München, 28(1977)4/5, S. 159-164.
Seidel, U. (1973): Ergebnisse und Perspektiven juristischer Benutzerforschung. In: NJW 1973, S. 1676ff.
Shebchenko, V. V. (1974): One approach to the problem of syntactic analysis. In: Kibern., Klev, SU (1974)4, S. 30-38.
Shingha1, R., Toussanit, G. (1979): A bottom-up and top-down approach to using context in text recognition. In: Int. j. of man-mach. stud., London, 11(1979)2, S. 201-212.
SIEMENS (1980): Automatische Selektion von Stichwörtern aus Texten. Passat BS2000 Verfahrensbeschreibung. München.
Siemens (1976): Verwendung der natürlichen Sprache im Dialogverkehr mit Informationssystemen. In: Datenverarbeitung, (1976), S.70-79.

Siemens (1977a): CONDOR. München.
Siemens (1977b): GOLEM/PASSAT (BS2000) . An jedem Arbeitsplatz schnell und kostengünstig informiert. München.
Siemens (1978): GOLEM (BS 2000). (GOLEM-Handbuch). München.
Silva, G. M. (1971): Center for information services, phase 2: detailed System design and programming. Phase 2a: final report, part 7: text processing. Los Angeles, Calif., USA.
Simitis, S. (1970): Informationskrise des Rechts und Datenverarbeitung. Karlsruhe.
Simitis, S. (1977): Bundesdatenschutzgesetz - Ende der Diskussion oder Neubeginn? In: NJW 17/1977, S. 729-737.
Simmons, R. F.: Natural Language Question Answering Systems. CACM 13 (1970) S. 1 5 ff.
Skelly, S. (1970): Computers and Statute Law. In: Law and Computer Technology. 2(1970), S. 30ff.
Soerensen, J. (1977): PRECIS als mehrsprachiges System. In: Kommission der Europäischen Gemeinschaften (1977): 3. Europäischer Kongress über Dokumentationssysteme und -netze, Luxemburg, 3.5.-6.5.1977, München, Bd. 1., S. 247-273.
Soerensen, J., Austin 0. (1976): PRECIS in a multilingual context - part 2. In: Libri, Kopenhagen, Dänemark, 26(1976)2, S. 108-139.
Sonderforschungsbereich 100 "Elektronische Sprachforschung". Projektbereich A (Germanistik). Teilprojekt Al (1978): Automatische Lemmatisierung. Universität des Saarlandes. Saarbrücken.
Sonderforschungsbereich 100, "Elektronische Sprachforschung", Projektbereich A (Hrsg.): SALEM. Ein Verfahren zur automatischen Lemmatisierung deutscher Texte. Tübingen 1980.
Sparck Jones, K. (1971): Automatic keyword classification for Information retrieval. Hamdon, Conn., USA.
Sparck Jones, K. (1973): Linguistics and information science. New York, London.
Sparck Jones, K. (1974): Automatic Indexing. In: Journal of Documentation 4, S. 393-432.
Stalcup, W. S., Petrarca, A. E. (1975): Automatic vocabulary control and its evaluation in computer-produced indexes. In: HUSBANDS 1975, S. 73-74.
Standera, O. (1971): COMPENDEX/TEXT-PAC. Retrospective search. Information Systems and Services division report 9 (ISSO report 9), Calgary, Kanada.
Steinmüller, W. (1970): EDV und Recht. In: JA-Sonderheft 6(1970).
Steinmüller, W. (Hrsg.) (1976): AVD und Recht. Einführung in die Rechtsinformatik und das Recht der Informationsverarbeitung. 2. völlig neu gest. und erw. Aufl., Juristische Arbeitsblätter - Sonderheft 6. Berlin.
Stevens, M. E. (1970): Automatic indexing. A state-of-the-art report. Washington, D.C., USA.
Stibic, V. (1977a): Remarks on the economic feasibility of automatic indexing. In: BATTEN 197.7, S. 130-133.
Stibic, V. (1977b): Documentretrieval met de computer. Mogelijkheden, economische effectiviteit en grenzen bij het gebruik van natuurlijke taal (Document retrieval by computer; the possibilities, economic effectiveness and limitations of the use of natural language.) In: Open, Den Haag, Niederland, 9(1977)11, S. 523-532.
Supper, R. (1978): Neuere Methoden der intellektuellen Indexierung. Britische Systeme unter besonderer Berücksichtigung von PRECIS. München, New York, London, Paris.
Svenonius, E., Schmierer, H. F. (1977): Current issues in the subject control of information. In: Libr. Quart., Chicago, I11., USA, 47(1977)3, S. 326-346.
Taeuber, D. (1978): CONDOR: Ein integriertes Datenbank- und Informationssystem. In: Nachr. f. Dokum., 29(1978)3, S. 127-130.
Thiel, M. (1977): Linguistische Verfahren in der Chemie-Dokumentation. In: Sprache und Da-

tenverarbeitung, 1(1977)2, S. 147-155.
Technische Hochschule Ilmenau, Institut für Informationswissenschaft, Erfindungswesen und Recht (INER), Ilmenau, DDR (1973): 8. Kolloquium über Information und Dokumentation. Themenkreis 3: Kompatibilität von Methoden und Geräten als Voraussetzung für Kooperation. Ilmenau, DDR, 28.11.-30.11.1973.
TELDOK (1978): TELDOK-Benutzerbeschreibung. Konstanz.
Tillmann, H. G. (1973): Linguistische Probleme der Datenverarbeitung im Recht. In: DSWR 1973, S. 290ff.
Uhlig, S. (1972/1973): Zur Problematik der Bewertung eines juristischen Informationssystems. In: DVR 1(1972/1973), S. 56ff.
Ungeheuer, G. (1977): Konzeption eines Informationssystems auf linguistischer Basis (ISLIB). In: Sprache und Datenverarbeitung, 1(1977), S. 46-53.
Universität Saarbrücken (1972): Sonderforschungsbereich elektronische Sprachforschung. Aspekte der automatichen Lemmatisierung. Saarbrücken.
Universität Bielefeld, Zentrum für Interdisziplinäre Forschung (1978): Methodological problems in automatic text processing. Bielefeld.
University of Alberta, Department of Computing Science (1972): Report on information retrieval and library automation studies. Edmonton, CA, USA.
Upton, C. C. (1971): Computerized communications citations technology. Paper prepared for a conference of the International Communication Association, Phoenix, Ariz., USA, 22.4.-24.4.1971. Athens, Ohio, USA.
Vaccary, E., Delaney, W., Chiesa, A. (1977): DARR: A free-text analysis system for the Automatic Documentation of Radiological Reports. In: Meth. of inform. in med., 16(1977)3, S. 144-153.
Wahrig, G. (1973): Anleitung zur grammatisch-semantischen Beschreibung lexikalischer Einheiten. Tübingen.
Wahrig, G. (1978): Deutsches Wörterbuch. Berlin, München, Wien.
Walker, D. E. (1973): Automated Language Processing. Annual eview of Information Science and Technology, Washington, D.C., USA.
Weber, H. J., Zimmermann, H. (1973) : Zur Verwertbarkeit der Großschreibung bei der automatischen Reduktion syntaktischer Wortformen-Mehrdeutigkeiten im Deutschen. In: Festschrift für Paul Grebe. Bd. 2, Düsseldorf.
Weber, H. (1976): Automatische Lemmatisierung. Sonderforschungsbereich Elektronische Sprachforschung, LA 15/1, Saarbrücken.
Wedekind, H. (1974): Datenbanksysteme I. Zürich.
Wedekind, H., Härder, T. (1976): Datenbanksysteme II. Zürich.
Weihermüller, M. (1976): Untersuchungen über Ranking-Algorithmen in Dokument-Retrieval-Systemen. In: MÜLLER 1976, S. 173-202.
Weinberg, B. H. (1975): Levels of linguistic analysis and Information processing. In: HUSBANDS 1975, S. 71-72.
Weisgerber, M. (1976): Automatische Konstituentenanalyse - Verbalgruppen. Linguistische Arbeiten 17, Saarbrücken.
Weissenborn, J. (1977): Zur Rolle und Form der Analyse in der maschinellen Übersetzung. In: Kommission der Europäischen Gemeinschaften: 3. europäischer Kongress über Dokumentationssysteme und -netze, Luxemburg, 03.05.-06.05.1977. Bd.1. München, S.536-554.
Werner, H. (1979): Benutzeranforderungen an JURIS. Beobachtungen bei JUDO. In: JURIS-Benutzertagung 1979.

Werner, H. (1979a): Erfahrungsbericht zur erster Phase der Thesaurusrelationierung. Bericht JUDO-B-13, masch., Regensburg.
Wersig, G., Neveling, U. (1975): Terminologie der Information und Dokumentation.
Wessel, A. E. (1975): Informationsretrieval und Automatisierung. Perspektiven und Probleme. Darmstadt.
Wieland, U. (1979): Recherche auf der Basis einer syntaxorientierten maschinellen Sprachanalyse. Diss. Regensburg.
Wilks, Y. (1974): One Small Head - Models and Theories in Linguistics. In: Found. of Lang., l1(1974), S. 77-95.
Wingert, F. (Hrsg.) (1978): Klartextverarbeitung. Berlin.
Winograd, T. (1972): Understanding Natural Language. Edinburgh.
Winograd, T. (1976): Artificial Intelligence and Language Comprehension. Washington, D.C., USA.
Woods, W. A., Kaplan, R. M., Nash-Webber, B. (1972): The Lunar Sciences Natural Language Information System: Final Report. Cambridge, Mass., USA.
Woods, W. A. (1973): An experimental parsing System for transition network grammars. In: RUSTIN 1973, S. 111-154.
Woldmann, K. (1971): Die Brauchbarkeit der Rangfolgekriterien für die automatische Textverdichtung. Abschlußarbeit, Technische Hochschule Ilmenau, Institut für Informationswissenschaft, Erfindungswesen und Recht (INER), Ilmenau, DDR.
Yu, C. T. (1973): A formal construction of term classes. Edmonton, Kanada.
Yu, C. T., Salton, G. (1975): The effectiveness of the thesaurus method in automatic information retrieval. Washington, D. C., USA.
Yu, C. T., Salton, G. (1976): Precision weighting - an effective automatic indexing method. In: J. of the ACM, Baltimore, Md., USA, 23(1976)1, S. 76-88.
Yu, C. T., Salton, G. (1977): Effectiv information retrieval using term accuracy. In: COMMUNICATIONS (USA), 20(1977)3,., S. 135-142.
Yu, C. T., Salton, G., Siu, M. K. (1978): Effective automatic indexing using term addition and deletion. In: J. of the ACM, New York, N.Y., USA, 25(1978)2, S. 210-225.
Zentralstelle für maschinelle Dokumentation (ZMD) (1977): Arbeitsbericht 1977. Frankfurt.
Zielinski, D. (1973): Das Juristische Informationssystem. In: DVR 2(1973), S. 36ff.
Zimmermann, H. (1972a): Zur Konzeption der automatischen Lemmatisierung von Texten. In: Universität Saarbrücken: Aspekte der automatischen Lemmatisierung, LA 12. Saarbrücken, S. 4-10.
Zimmermann, H. (1972b): Das Lexikonsystem zur maschinellen Sprachbearbeitung. In: Universität Saarbrücken: Aspekte der automatischen Lemmatisierung, LA 12. Saarbrücken, S. 62-81.
Zimmermann, H. (1972c): Das Lexikon in der maschinellen Sprachanalyse. Frankfurt/M.
Zimmermann, H. (1974): Ein Konzept zur syntaktischen Oberflächenanalyse. In: IRAL-Sonderband: Kongreßbericht der 5. Jahrestagung der Gesellschaft für Angewandte Linguistik, Heidelberg 1974, S. 172-179.
Zimmermann, H. (1978a): Probleme der automatischen Indexierung von Fachtexten am Beispiel juristischer Dokumente. In: WINGERT 1978, S. 112-121.
Zimmermann, H. (1978b): Automatische Textanalyse und Indexierung. LDV-Aktivitäten in Regensburg - 1. Teil. In: KRALLMANN 1978, S. 20-33.
Zimmermann, H. (1979a): Ansätze einer realistischen automatischen Indexierung unter Verwendung linguistischer Verfahren. In: KUHLEN 1979, S. 311-338.
Zimmermann, H. (1979b): JUDO - Modell einer computergestützten Indexierung auf linguisti-

scher Grundlage. (JUDO-V-05), masch., Regensburg.
Zimmermann, H.: Das Projekt JUDO-DS - Juristische Dokumentanalyse im Bereich Datenschutz. In: Das Inforum (1981), S. 50-65.
Zimmermann, H. (1982): Automatic Indexing and Retrieval as a tool to improve information and technology transfer In: Proceedings of the Lind scientific meeting COMPUTER PROCESSING OF LINGUISTIC DATA: Bled (Jugoslawien), 7.-9. Oktober 1982.
Zur Sache 5/74: Datenschutz/Meldegesetz. Sachverständigenanhörung, Gesetztestexte. Presse- und Informationszentrum des Deutschen Bundestages (Hrsg.).