In: Frank W. Felzmann et al. (1980): JUDO - Juristische Dokumentanalyse. Universität Regensburg: Abt. für Nichtnumerische Datenverarbeitung. S. 22-54

Harald H. Zimmermann

1 JUDO - Zielsetzung, Vorgehensweise und erreichte Ergebnisse

1.1 Einordnung in den Themenkreis der Informationswissenschaft

Im folgenden soll ein Forschungsprojekt vorgestellt werden, das Fragen der computergestützten Erschließung (Indexierung, Deskribierung) von Fachtexten und darauf aufbauende Retrievalverfahren (Referenz-Retrieval) in den Mittelpunkt stellt. Ehe jedoch auf die diesbezüglichen konkreten Ziele, Verfahrensweisen und die bisher vorliegenden Ergebnisse eingegangen wird, soll der Versuch unternommen werden, dieses spezifische Vorhaben in den Gesamtzusammenhang der Informationswissenschaft zu stellen.

Insbesondere anhand der 1979 ergangenen Empfehlungen des Sachverständigenkreises 'Ausbildung im IUD-Bereich' des Bundesministerium für Forschung und Technologie (BMFT) mit dem Ziel der "Förderung informationswissenschaftlicher Forschung an wissenschaftlichen Hochschulen" (AUSBILDUNG 1979) ist die Möglichkeit gegeben, die im Titel angesprochenen Fragestellungen in die Forschungs- und Ausbildungsbereiche der Informationswissenschaft einzubinden. Zugleich wird dabei deutlich, dass die vorzustellenden Probleme nicht nur unter einer einzigen Thematik relevant erscheinen. So soll die informationswissenschaftliche Forschung und Entwicklung z.B. unter dem Aspekt der Fach-Kommunikations- und -Informationsprozesse die Wirksamkeit von Informationseinrichtungen verbessern helfen. Ausgangspunkt ist dabei die gegenwärtige Praxis der Fachinformation und -kommunikation, v.a. die Kommunikation über fachliche Inhalte in Verwaltung, Wissenschaft, Wirtschaft und Gesellschaft (vgl. AUSBILDUNG 1979, S. 2).

Unter diesem Schwerpunkt sollen mit Bezug auf Fragestellungen, die sich Begriffen wie 'Indexierung', 'Referierung', 'Retrievalverfahren', 'Suchstrategien', 'Relevanzkontrolle' (vgl. AUSBILDUNG 1979, S. 6) zuordnen lassen, explizit - aus der Sicht des Endbenutzers, also z.B. des Fachwissenschaftlers - verbesserte Verfahren zur Informationsermittlung, -vermittlung und -verbreitung gefördert werden.

Besonderes Augenmerk gilt dabei den "bei Analyse, Synthese und Verbreitung von relevanten Daten auftretenden linguistischen und kognitiven Problemen" (AUSBILDUNG 1979, S. 8), die unter dem Aspekt der "Repräsentation und Transformation von Wissen" einem besonderen Forschungsschwerpunkt zugeordnet wurden. Zu den Verfahrensweisen werden wiederum explizit "computergestützte Sprachanalyse, -übersetzung und -Synthese" sowie die "Sprache in Informationsprozessen" selbst gerechnet.

Es ist nicht schwer, die Themenstellung des hier vorgestellten Forschungsvorhabens daneben auch weiteren Schwerpunkten informationswissenschaftlicher Forschung, Entwicklung und Ausbildung zuzuordnen; insbesondere spielen allgemeine soziologische Fragen (z.B. Benutzerzugang/ Barriereforschung) hierbei eine Rolle.

In der gegenwärtigen Forschungsförderung im Rahmen des Programms der Bundesregierung zur Förderung von Information und Dokumentation (IuD-Programm) wird den Fragen der "Repräsentation und Transformation von sprachlich vermitteltem Wissen" (vgl. AUSBILDUNG 1979, S. 8) besondere Beachtung geschenkt. Dies verdeutlicht auch die Einrichtung eines speziellen Förderschwerpunkts zur Informationslinguistik. In einer Empfehlung zu diesem Bereich wird vertiefend auch auf Fragestellungen Bezug genommen, wie sie im Mittelpunkt des vorzustellenden Forschungsvorhabens stehen (vgl. INFORMATIONSLINGUISTIK 1979). Hierzu gehört z.B. die "Entwicklung von höheren Dokumentationssprachen", wobei insbesondere auch das Verhältnis von "Dokumentationssprachen zu anderen Sprachen, wie natürlichen Sprachen, Fachsprachen" angesprochen ist.

Zum Themenkreis der "Verbesserung und Automatisierung von Inhaltserschließungsverfahren" wird unter anderem das "Klassifizieren, Indexieren, Referieren und Verdichten von Texten und der Aufbau von Wörterbüchern und Thesauri unter Anwendung linguistischer und statistischer Verfahren" (INFORMATIONSLINGUISTIK 1979, S. 3) gerechnet.

Neben einer Reihe weiterer natürlichsprachig orientierter Themenbereiche ist im vorliegenden Kontext die Frage der "Erweiterung des Dokument-Referenz-Retrieval" von besonderem Gewicht. Mit den Themen "Verbesserung der Qualität der Erschließungskomponenten", "Einsatzmöglichkeiten linguistischer Verfahren zur Verbesserung der Leistung gegenwärtiger Retrieval-Systeme", "Entwicklung benutzerfreundlicher Anfragesprachen" (INFORMATIONS-LINGU-ISTIK 1979, S. 4) ist gerade ein Bereich angesprochen, der im Zentrum des vorzustellenden Forschungsvorhabens steht.

1.2 Erläuterung der Themenstellung und Ziele

1.2.1 Begriffsklärungen

Das Forschungsvorhaben hat die 'Modellentwicklung eines Verfahrens zur computergestützten Indexierung am Beispiel der Analyse juristischer Dokumente' zum Ziel (das Akronym JUDO steht verkürzend für Juristische Dokumentanalyse).

- Unter Modellentwicklung soll dabei nicht allein eine theoretische Konzeption, sondern die Entwicklung eines Laborsystems verstanden werden, in dessen Rahmen exemplarisch die entsprechenden Fragen abgehandelt werden. Zugleich wird der Anspruch der Übertragbarkeit des Systems auf andere als die behandelten Rechtsgebiete, z.B. auch auf andere Fachgebiete erhoben.

- Unter Indexierung wird - in Anlehnung an WERSIG/NEVELING 1975 - die Zuweisung von Termini (engl. 'index terms') zu Dokumenten oder Objekten verstanden, die einen Informationssuchenden in die Lage versetzen, diese Dokumente oder Objekte aufgrund von ausgewählten (eindeutigen) Begriffen (concepts), die mit diesen Termini verbunden sind, wiederzufinden (zu 'retrieven'). Der Begriff 'Deskriptor' ist für den vorhandenen Zweck im Folgenden mit 'Terminus' gleichgesetzt, obgleich 'Deskriptor' in der Literatur häufig in eingeschränkter Form verwendet wird.

  Nicht berücksichtigt wird zunächst die Frage einer Selektion der sinntragenden Termini

im Hinblick auf ihre Relevanz zur Deskribierung eines aktuellen Dokuments, während es prinzipiell vorgesehen ist, (aus fachlicher Sicht) bei Deskriptoren generelle Relevanzangaben zu vermerken und bei der Deskriptorenvergabe zur Selektion auszunutzen. Die Frage der Gewichtung von Deskriptoren, wie sie bereits implizit durch Vergabe bzw. Nichtvergabe von in einem Dokument vorhandenen Begriffen auch bei der intellektuellen Deskribierung vorgenommen wird, ist im Rahmen der darzustellenden Projektphase also ausgeklammert worden; es wird auch abgesehen von echter 'Gewichtung' aufgrund der Relevanz eines Deskriptors für die Beschreibung eines Dokuments. Nur am Rande behandelt ist auch das Problem der Vergabe von Deskriptoren (Schlagwörtern), die nicht explizit im Dokumenttext verzeichnet sind: hier kann gegenwärtig in begrenztem Umfang über Thesaurusrelationen eine Zuordnung erfolgen.

Es ist also festzuhalten, dass im Rahmen von JUDO eine Einschränkung der Indexierung auf die Verarbeitung textueller Informationen erfolgt und dass innerhalb dieser Verarbeitung derzeit keine weitere Selektion oder Gewichtung der ermittelten Deskriptoren erfolgt. Diese Beschränkung ist besonders wichtig im Hinblick auf die Interpretation der in der Dokumentation üblicherweise zur Bewertung der Qualität von Indexierungs- und Retrieval-Verfahren verwendeten Messwerte 'Recall' und 'Precision', die auch für die Beurteilung der Leistungen des JUDO-Systems von Bedeutung sind. Unter 'Recall' wird allgemein das Verhältnis der gefundenen relevanten Dokumente zu den in der Informationsbank vorhandenen relevanten Dokumenten verstanden. Im günstigsten Fall werden alle relevanten Dokumente aufgrund eines Suchvorgangs gefunden.

Unter 'Precision' wird üblicherweise das Verhältnis der gefundenen relevanten Dokumente zu den bei dem Suchvorgang (überhaupt) gefundenen Dokumenten betrachtet. Je weniger relevante Dokumente unter den gefundenen vorhanden sind, desto niedriger ist die Precision. Im Idealfall sind alle gefundenen Dokumente relevant.

Die Beurteilung der Relevanz (d.h. des Zutreffens) von Dokumenten im Hinblick auf eine Suchanfrage ist in der Praxis oft subjektiv. Dieser Maßstab lässt sich jedoch bei Nicht-Berücksichtigung der Gewichtung von Deskriptoren vereinfachen und 'objektivieren', wenn er auf die Identifikation eines Begriffs (Deskriptors bzw. Deskriptorkomplexes) übertragen wird.

Im Rahmen des JUDO-Konzepts bedeutet Relevanz die korrekte maschinelle Identifikation eines Deskriptors, d.h. das korrekte Feststellen des Vorkommens eines Begriffs (nach Form und Bedeutung) in einem Dokument. Recall ist somit das Verhältnis der gefundenen Dokumente, in denen ein bestimmter Begriff ('type') identifiziert wurde, zu den Dokumenten, in denen der Begriff in einer Ausprägung ('token') auftritt. Im günstigsten Fall werden alle Dokumente gefunden, in denen eine graphische Repräsentation (Wort, Wortform, Schreib- und Bedeutungsvariante) dieses Begriffs auftritt.

Precision ist analog dazu das Verhältnis der Anzahl der gefundenen Dokumente, in denen ein bestimmter Begriff (type) korrekt identifiziert wurde, zu der Anzahl der Dokumente, die insgesamt gefunden (nachgewiesen) werden. Fehler sind dabei zurückzuführen auf falsche Deskriptor-Zuordnungen bei der linguistischen Analyse.

Unter dieser Einschränkung werden im vorliegenden Fall die Termini 'Recall' und 'Preci-

sion' verwendet. (Vgl. dazu auch KUHLEN 1977, S. 157ff. Allerdings bezieht sich Kuhlen bei der "Zusammenführung" auf die korrekte Zuordnung von Wörtern zu Stämmen bzw. Grundformen, während hier die Zuordnung zu Deskriptoren - auch unter semantischer Vereindeutigung - im Mittelpunkt steht. Methodisch gesehen gelten jedoch die gleichen Voraussetzungen bzw. Einschränkungen).

- 'Computergestützt' heißt das System, weil maschinelle Verfahren den Prozess der intellektuellen Indexierung in wesentlichen Verarbeitungsschritten unterstützen sollen. Der Begriff 'automatisch' wird vermieden, da nicht der Eindruck erweckt werden soll, das gesamte Verfahren könne 'vollautomatisch', also z.B. auch ohne die intellektuelle Pflege von Lexika und Thesauri, ablaufen, wenn auch z.Zt. keine intellektuelle 'Nachredaktion' der automatisch gewonnenen Deskriptoren vorgesehen ist. Damit steht die Vorgehensweise in gewissem Gegensatz zu dem derzeit ebenfalls in einer Laborversion verfügbaren System CONDOR (vgl. BANERJEE 1977), das - ausgehend von einem automatischen Sprachanalyseverfahren - bei der Produktion von Dokumentanalysen ohne intellektuelle Aufbereitung der Daten auskommen will.

- Die Grundlage von JUDO heißt linguistisch, weil morphologische, syntaktische und semantische Sprachanalysemethoden zum Einsatz kommen. Der Begriff 'linguistische Verfahren' muss hier also weiter gefasst werden als z.B. bei dem Siemens-Software-Produkt PASSAT (vgl. HOFFMANN 1971) oder bei STAIRS-TLS von IBM. PASSAT (während der Deskribierung) und STAIRS-TLS (während des Retrieval) ordnen den aus dem laufenden Text isolierten Wortformen die möglichen Grundformen zu, die als Deskriptoren oder Zugriffswörter verwendet werden können; die Verfahren sind jedoch auf kontextfreie (morphologische) Einzelwortanalysen beschränkt. In JUDO werden dagegen eine Syntaxanalyse und in Ansätzen semantische Verfahren verwendet.

- Als Juristische Dokumente wurden während des Aufbaus des JUDO-Systems hauptsächlich Gesetze und Gesetzesentwürfe zum Datenschutz verwendet. Die Verfahren sind jedoch ebenso für Judikate (d.h. Urteile), Literatur (z.B. Abstracts), aber auch Berichte, Verordnungen usw. bis hin zu ('halbfachlichen') Zeitungsmeldungen anwendbar. In jedem Falle steht die Bearbeitung des Freitexts (Klartexts) im Mittelpunkt. Fragen des strukturierten Überbaus der Dokumente (bei Normen z.B. Datum des In-Kraft-Tretens) werden bei der (vor)gegebenen Fragestellung nicht explizit behandelt, wohl aber werden wenigstens partiell derartige Informationen in das Retrievalsystem integriert, um eine angemessene Benutzersituation zu schaffen, die gerade durch die Kombination von formatierten mit sog. unformatierten (d.h. textuellen) Daten gekennzeichnet ist.

1.2.2 Zielsetzung

In der jüngsten Zeit werden in der Bundesrepublik Deutschland erste Forschungsvorhaben zu Sprachverstehensprozessen mithilfe des Computers (einem Teilbereich der 'Artificial Intelligence') durchgeführt (vgl. z.B. das Vorhaben HAM-RPM in Hamburg; s. HAHN 1978). Sie bilden insofern einen gewissen Kontrast zu dem vorzustellenden System, als sie meist sehr kleine Sprach- und Weltausschnitte behandeln, allerdings unter Verwertung weitergehender (zum Teil enzyklopädischer) Informationen, z.B. mit dem Ziel einer Simulation menschlichen Sprachverstehens und -verhaltens.

Während solche Verfahren - für die eingeschränkten (Teil-)Welten - durchaus in der Lage zu sein scheinen, präzisere Informationen über die Miniwelt aus sprachlichen Äußerungen zu erschließen, sind sie bislang nicht auf größere Sprach- und Dokumentationsbereiche übertragbar und erscheinen aufgrund des erforderlichen Aufwands wenig kostengünstig; sie sind im wesentlichen noch der Grundlagenforschung zuzuordnen und allenfalls in Laborsituationen einsetzbar. Oft sind in dem möglichen (eingeschränkten) Einsatzbereich der Artificial-Intelligence-Verfahren die Problem- und Fragestellungen vom Ansatz her begrenzt, so daß die natürlichsprachige Kommunikation in Konkurrenz zu Verfahren wie der 'Menütechnik' oder einem 'graphischen Dialog' steht, wobei diese Alternativen häufig den Vorteil einer präziseren und rascheren Verarbeitung aufweisen. Dagegen - zumindest ist dies die Vorstellung, die der Konzeption von JUDO zugrundeliegt - ist ein erheblicher Bedarf an Verfahren gegeben, die in dem Bereich der Dokumentation von textuellen Massendaten gegenüber vorhandenen Verfahren verbesserte Resultate bringen; Beispiele für Bereiche mit möglichen großen Textdatenvolumen sind besonders die Dokumentation von Normen und Judikaten (mehrere tausend Dokumente/Jahr) im Rechtswesen (vgl. JURIS bzw. die Steuerrechtsdokumentation der DATEV), die Befunddokumentation in der Medizin, schließlich die Textdokumentation im Büro oder in der Verwaltung.

Die Referenz-Dokumentation - traditionell so genannt, weil nicht die unmittelbaren, im Text kodierten Fakten, sondern nur mittelbare Verweise oder 'Referenzen' zu den präzisen Fakten das Bindeglied zwischen Dokumenterschließung und Problemlösung beim Retrieval darstellen - ist das heute allgemein übliche Verfahren der Dokumentation. In diesem Bereich sind verschiedene Systeme im praktischen Einsatz, die auch textuelle Informationen verarbeiten und teilweise erschließen. Zwei der bekannteren Verfahren seien kurz erwähnt, da JUDO sich zunächst an ihnen ausrichtet:

Mit dem GOLEM-Teilsystem PASSAT (vgl. HOFFMANN 1971) ist es möglich, (Einzel-)Wortformen eines Textes (Dokuments) anhand einer sog. Vergleichswortliste (VWL) zu identifizieren. Die VWL enthält u.a. nicht-zusammengesetzte Wörter, Endungen und Fugenlisten. Aufgrund dieser Angaben kann in der Regel eine Reduktion der Textwortformen auf die mögliche(n) Grundform(en) erfolgen. Zusammengesetzte Wörter (Komposita) werden als solche ermittelt und in ihre (möglichen) Bestandteile zerlegt, wenn diese Teile in der Vergleichswortliste (VWL) vorhanden sind. Invariante mehrwortige Ausdrücke können als solche erkannt und zu einem Begriff zusammengefasst werden, wenn sie kontinuierlich, d.h. unmittelbar hintereinander auftreten. Die (meist intellektuell gepflegte) Assoziationsmatrix wird im Wesentlichen zur Eliminierung von STOP-Wörtern verwendet.

Obwohl bei der Pflege der Vergleichswortliste (VWL) bereits ein beachtlicher Aufwand nötig ist (Aufnahme aller nicht zusammengesetzten Wörter und Aufbau von Endungs- und Fugenlisten), lassen sich sprachlich falsche Zuordnungen oder Zerlegungen bei der Dokumenterschließung nicht völlig vermeiden; das im Deutschen nicht unbeträchtliche Problem der mehrdeutigen Wörter (Homographen und Homonyme) und damit der Bedeutungsdifferenzierung ist für den Einzelbeleg im Dokument nicht zu lösen. Dabei hat sich herausgestellt, dass ca. 15 % der in den im Rahmen des JUDO-Projekts untersuchten Fachtexten (zum Datenschutzrecht) vorkommenden Substantive und rd. 20 % der Verben zumindest allgemeinsprachlich gesehen semantisch mehrdeutig sind. Selbst wenn man davon ausgeht, dass komplexe Suchanfragen (etwa bei der Verknüpfung mehrerer Deskriptoren mit 'logischem UND' oder die Verwendung der GOLEM-Feinrecherche) einen Teil der Probleme bei einer fehlenden Disambiguierung vermeiden können,

bleibt die Frage der Akzeptanz dieser Retrievaltechnik durch den Benutzer, der sich solchen für sein Sprach- und Sachverständnis unnatürlichen Vorgängen beim Retrieval (zur Vermeidung von Ballast) gegenübersieht.

Das zweite hier beispielhaft zu erwähnende Verfahren, das System STAIRS, erlaubt - ähnlich wie PASSAT, wenn auch über andere Verfahrensweisen - ebenfalls das Retrieval über Grundformen (in der Variante STAIRS-TLS); mithilfe von sog. Wortmaskierungen (Truncation) lassen sich bei STAIRS auch Teilworte (etwa unter Vernachlässigung von Wortendungen) recherchieren. Aber auch hier sind einige natürlichsprachliche Probleme (z.B. Auflösung syntaktischer/semantischer Mehrdeutigkeiten, abtrennbare Wortbestandteile) nicht gelöst, einige Fragen (z.B. der Identifikation von Flexionsformen in Kompositabestandteilen) noch offen.

Bei beiden an dieser Stelle nur in Grundzügen (und damit sicherlich etwas vergröbernd) aufgezeigten Verfahren können allenfalls - und das nur unvollständig und mit 'technischer' Belastung des Benutzers - die 'gängigen' flexionsmorphologischen Probleme (Abbildung kontinuierlicher Textwortformen auf Grundformen) - als gelöst angesehen werden; das Problem der Auflösung von Wortmehrdeutigkeiten wird in gewisser Weise 'überspielt' (und somit nur unvollständig gelöst) durch Positionsangaben zu Kontextbegriffen (z.B. A 'im gleichen Satz wie' B; A 'im gleichen Abschnitt wie' B; A 'neben' B); eine korrekte Identifikation der Grundformen kann bei diskontinuierlichen (abtrennbaren) Wortbestandteilen nicht geleistet werden, die Ermittlung von Deskriptoren durch den Algorithmus zur Kompositazerlegung führt stets zu einer gewissen Fehlerquote.

Vor diesem Hintergrund und im Kontrast dazu war zunächst die Konzeption von JUDO entstanden. Diese Konzeption lässt sich anhand von drei Komponenten wie folgt beschreiben:

(1) die Indexierung und letztlich auch das Retrieval erfolgen unter Verwendung linguistischer Verfahren;

(2) das System wird an Fachtexten - hier zunächst Rechtstexten aus dem Bereich Datenschutz - in einem Laborsystem praktisch erprobt;

(3) bestehende Software findet nach Möglichkeit Verwendung.

Obgleich also ein nicht-statistischer Ansatz im Mittelpunkt der Inhaltserschließung steht, werden statistisch-probabilistische Verfahren integriert, ohne dass dazu eigene intensive Forschungen ausgeführt werden. Im wesentlichen soll versucht werden, den Nachweis zu führen, dass linguistische Verfahren im Rahmen eines Indexierungs- und Retrievalsystems für textuelle Daten sinnvoll einsetzbar sind.

Es kann im Rahmen der Modellentwicklung zunächst nicht erwartet werden, dass dieser Nachweis über einen konkreten Anwendereinsatz geführt wird. Allerdings sollen die prinzipiellen Möglichkeiten und Vorteile, auch die Operationalität und Funktionalität der Verfahren unter Beweis gestellt werden. Im Verlaufe der zweiten Projektphase (1980/1981: JUDO-DS) sollen dazu erste Evaluierungen vorgenommen werden, auch um möglicherweise eine ausreichende Entscheidungsbasis für spätere (Pilot-)Anwendungen zu schaffen.

Obwohl - in der Modellversion von JUDO - eine ökonomische Betrachtung nur in Ansätzen vor-

genommen werden kann, wird sich dieses Verfahren daher mittelfristig nicht nur im Hinblick auf die Qualität der Ergebnisse, sondern auch in Bezug auf Kosten-Nutzen-Fragen mit den am Markt befindlichen Systemen messen lassen müssen. Dennoch wird bei der folgenden Betrachtung der qualitative Aspekt im Vordergrund stehen (zur Gesamtübersicht der JUDO-Komponenten vgl. Abb. 1.1)

Die Reduktion von Textwortformen auf Grundformen ist im Rahmen von JUDO ein (notwendiger) Teilschritt. Entscheidend ist dabei jedoch die Einbeziehung syntakto-semantischer Analysen. Dadurch soll sichergestellt werden, dass eine Wortform nicht - wie bei PASSAT - auf alle potentiellen Grundformen reduziert wird, sondern auf die im jeweiligen Kontext aktuelle (d.h. zutreffende) Grundform. Z.B. soll die Zeichenfolge ABSCHNITTEN - je nach Kontext - auf das Verb ABSCHNEIDEN oder das Nomen ABSCHNITT zurückgeführt werden. Daneben sind semantische Mehrdeutigkeiten aufzulösen (ANLAGE ist (u.a.) ggf. zu identifizieren als 'Rechenanlage' oder als 'Material'/'Papier' usf.). Bei der morphologischen Analyse sind eine Reihe von Detailfragen systematisch zu behandeln und zu bewältigen, z.B. das Umlautproblem (LAENDER -- LAND; FAELLE -- FALL/FAELLEN) und sonstige sprachliche 'Unregelmäßigkeiten' (z.B. Ablaut: GING -- GEHEN).

Als kompliziert erweist sich in diesem Zusammenhang die Identifikation sog. diskontinuierlicher Einheiten, wie etwa BESTEHEN ... FORT -- FORTBESTEHEN. Daneben sind - auch hier über die erwähnten Verfahren PASSAT und STAIRS/TLS hinausgehend - 'Feste Wendungen' im linguistischen und fachlichen Sinne unmittelbar als Deskriptoren verfügbar zu machen (Beispiel: BUNDESUNMITTELBARE KOERPERSCHAFT, PERSONENBEZOGENE DATEN, IN KRAFT TRETEN, RECHT AUF AUSKUNFT). Dies soll zugleich veranschaulichen, dass mit der Verwendung natürlicher Begriffsformen (Deskriptoren) den (sprachlich-fachlichen) Gewohnheiten. des Benutzers entgegengekommen werden soll.

Während des Projektverlaufs wurde zunehmend deutlich (vor allem zeigte sich dies an den ersten untersuchten Texten zum Bereich Datenschutz), dass eine maschinelle Sprachanalyse über die Auswertung der ermittelten Sprach- oder (derzeit) Satzstrukturen für die Indexierung hinaus weitere Möglichkeiten bietet, den Retrievalvorgang im Sinne einer Vereinfachung und Erhöhung der Effizienz zu unterstützen. Es war bislang nicht möglich, alle sich dabei bietenden Möglichkeiten zur Auswertung der Sprachanalyseergebnisse 'umzusetzen'. Eine vorläufige Untersuchung der von (juristischen) Experten anhand von Fachlexika und Registern zu Kommentaren ermittelten mehrwortigen Begriffe (insbesondere der sog. 'Festen Wendungen') ergab jedoch, dass diese zusammengesetzten Begriffe spezifische syntaktische Oberflächenstrukturen aufweisen, wie sie auch von der maschinellen Analyse als Teilstrukturen gewonnen werden:

Beispiele: (Fachbegriff)		(syntaktische Struktur)
STRAFBARE HANDLUNG; SCHUTZWÜRDIGE BELANGE	--	Adjektivattribut + Nomen
RECHT AUF SPERRUNG; GESETZ ÜBER DAS MELDEWESEN	--	Nomen + Präpositionalattribut
ZUSTIMMUNG DER BETROFFENEN;		

FORM DER AUSKUNFTSERTEILUNG -- Nomen + Genitivattribut

Da den Fachwissenschaftlern keine feste Anleitung gegeben worden war (und auch in der Praxis kaum gegeben werden kann), welche der Deskriptoren unmittelbar als 'lexikalisierte' mehrwortige Einheiten betrachtet werden sollten, ergab sich - zumal bei einem noch relativ jungen Fachgebiet wie dem Datenschutzrecht - eine Reihe von Zweifelsfällen. Das Problem lässt sich jedoch zumindest eingrenzen, wenn die Möglichkeit eröffnet wird, durch die Satzanalyse ermittelte syntaktische (Teil-)Strukturen in die Deskribierung und damit auch das Retrieval einzubeziehen.

Zunächst werden daher alle syntaktischen Relationen der Form

Adjektivattribut - Nomen	(Relator: A)
Nomen - Genitivattribut	(Relator: G)
Nomen - Präpositionalattribut	(Relator: P)
Nomen - nominale Anreihung	(Relator: K)

als 'Komplexe Deskriptoren' unter Markierung des erforderlichen Relators (A, G, K, P) zum Retrieval bereitgestellt, soweit sie über die maschinelle Analyse als solche erkannt werden. Diese Struktur wird auch dann recherchierbar, wenn aufgrund des semantischen Lexikons eine 'Feste Wendung' erkannt wurde. Eine entsprechende intellektuell vergebene Synonymierelation im Thesaurus (vgl. Kap. 1.5 und Kap. 4.2.4.2.12) sorgt dafür, dass der Benutzer in diesen Fällen mit einer beliebigen Variante (Feste Wendung bzw. Komplexer Deskriptor) recherchieren kann. Der Systemexperte (als Informationsvermittler) bzw. der Fachmann wird eher die bequemere Form der 'Festen Wendung' bevorzugen; im Falle eines Fehlschlagens beim Retrieval empfiehlt sich die relationierte Form. Natürlich ist es denkbar, über die derzeit aufgebauten 'Minisequenzen' hinauszugehen und auch komplexere (Nominal-)Strukturen anzubieten bzw. auszuwerten, sei es zur Indexierung oder auch zur Textextraktion (vgl. dazu BRAUN 1973, SEELBACH 1975 und ROSTEK 1979); solche Überlegungen wurden jedoch bislang aus Zeitgründen zurückgestellt.

1.3 Verfahrensschritte zur Sprachanalyse

1.3.1 Grundlagen

Im JUDO-System wird das an der Universität des Saarlandes in Entwicklung befindliche Sprachanalysesystem SATAN (Saarbrücker Automatische Text-ANalyse) zugrundegelegt, da dieses Verfahren als gegenwärtig einziges System (neben dem nicht-lexikalischen Verfahren CONDOR) in der Lage ist, deutschsprachige Texte (Sätze) relativ uneingeschränkter Struktur mit für die Referenz-Dokumentation ausreichender Sprachanalysetiefe zu verarbeiten. SATAN bietet zur Zeit genau jene Komponenten an, die für die in JUDO vorgestellten Zielsetzungen relevant erscheinen (historisch gesehen waren gerade die erkennbaren Möglichkeiten von SATAN auch der Ausgangspunkt für die Konzeption von JUDO). Dazu gehören:

- Reduktion von Wortformen auf Grundformen,
- Disambiguierung von Homographen/Homonymen,
- Aufbau syntaktischer Strukturen (z.B. Erkennung von Nominalgruppen und nominalen Subgruppen, Verbalgruppenanalyse, Haupt-, Nebensatzzuordnungen).

Zugleich erklärt die Verfahrensweise von SATAN, warum bei JUDO eine vollständige syntaktische Sprachanalyse vorausgesetzt wird, um scheinbar trivialere (Teil-)Strukturen wie Nominal-

gruppen zu identifizieren: Vereinfacht ausgedrückt benötigt SATAN eine komplette Identifikation mindestens der (syntaktischen) Satzstruktur, um unter den möglichen 'Lesarten' der Oberflächenwörter (und den potentiellen Teilstrukturen) relativ zuverlässig die aktuelle Form - d.h. z.B. die im vorliegenden Text gegebene oder zumindest mögliche Struktur - zu ermitteln. Sofern die Verfahrensschritte von SATAN auch Teilschritte von JUDO darstellen, sind sie in der folgenden Beschreibung implizit enthalten. Obwohl also - wie die Arbeiten von BRAUN 1973 und v.a. ROSTEK 1979 zeigen - ein partielles Parsing durchaus praktikable Ergebnisse bringen kann, sind einer solchen Vorgehensweise Grenzen gesetzt (bedingt durch syntaktische Homographie, nominale und verbale Einbettungen bzw. Diskontinuitäten), die nur durch ein vollständigeres Parsing überwunden werden können. Allerdings ist auch im JUDO-System ein partielles Parsing für den Einzelfall vorgesehen, dass eine tiefergehende Analyse (eines Satzes) nicht erfolgreich durchgeführt werden konnte.

1.3.2 Lexikonkomponente

JUDO ist - wie SATAN und im Gegensatz zu CONDOR - gekennzeichnet von einer starken lexikalischen Komponente. Normalerweise wird bei der maschinellen Komponente von JUDO vorausgesetzt, dass zu allen Textwortformen Einträge in den verschiedenen Lexika vorliegen. Es wird dabei derzeit unterschieden zwischen einer morphologisch-syntaktischen Lexikonkomponente und einem semantischen Regel-Lexikon.

Beide Lexika stellen im Wesentlichen Stammformenlexika dar; nur bei den sog. Funktionswörtern (z.B. Artikel, Pronomen, Konjunktionen, Präpositionen sowie den unflektierten Adverbien) werden Wortformeneinträge zugrundegelegt.

Die Funktionswörter - z.Zt. ca. 2.000 Einträge - stellen i.a. allgemeinsprachliche Einheiten dar, d.h. sie sind in Inhalt und Funktion fachgebiets- und textsortenunabhängig. Es kann daher vorausgesetzt werden, dass die Pflege dieser Einträge nicht vom Systemanwender zu übernehmen ist. Dies gilt bereits derzeit für JUDO, wo die (inzwischen zu vernachlässigenden) diesbezüglichen 'Lücken' im Lexikon durch die Saarbrücker Projektgruppe (d.h. die Entwicklungsgruppe von SATAN) geschlossen werden.

Anders sieht es aus im Bereich der 'Vollwörter', zu denen in erster Linie die Substantive, Verben und Adjektive gerechnet werden. Zu den Informationen, die im morpho-syntaktischen Lexikon kodiert werden müssen, sind zu rechnen: Deklinations- und Konjugationshinweise, syntaktische (potentielle) Valenzangaben (z.B. zu Oberflächensubjekt/-objekt), Angaben zu Nebensatzanschlüssen. Prinzipiell ist es hier vorstellbar, im Rahmen eines Systems, das über ausreichende technische Kapazitäten (z.B. Plattenspeicher großer Dimension mit schnellem Zugriff) verfügt, den Erstellungsaufwand für das morpho-syntaktische Lexikon längerfristig durch Bereitstellung eines größeren allgemeinsprachlichen Vokabulars zu reduzieren, wobei dieses mindestens den Umfang der 'gängigen' Wörterbücher (Wahrig, DUDEN, Langenscheidt) haben, also ca. 80.000 - 100.000 Grundformen identifizieren sollte.

Man kann sich auch vorstellen, dass durch fortlaufendes Ergänzen des Vokabulars in Abhängigkeit von den verarbeiteten (Fach-) Texten langsam (erfahrungsgemäß zunächst sehr langsam) eine Sättigung eintritt, so dass der Kodieraufwand für den Systemanwender - zumindest für das morpho-syntaktische Lexikon - im Laufe der Zeit erheblich reduziert werden kann. Diese Situa-

tion ist derzeit weder bei SATAN noch bei JUDO erreicht; beide liegen als Laborversionen vor (das syntaktische Lexikon umfasst derzeit ca. 20.000 Einträge), so dass bei einer Erweiterung der Textbasis noch ein beträchtlicher Aufwand bei der Lexikonkodierung betrieben werden muss.

Gerade fürs Deutsche (man denke v.a. an die Kompositabildung) ist eine streng lexikalistische Vorgehensweise in dem Sinne, dass ein Satz oder Text nur dann maschinell verarbeitet werden kann, wenn seine Wörter alle lexikalisch identifiziert werden konnten, natürlich unzureichend. Das System SATAN erlaubt daher u.a..

- eine automatische Kompositaerkennung und -zerlegung, vorausgesetzt, alle Teile sind (z.B. über vorhandene Simplizia) lexikalisch identifizierbar. Das Verfahren ist recht brauchbar, allerdings können bei diesem Erkennungsverfahren wieder Zerlegungsfehler (ARBEIT-SAMT vs. ARBEITS-AMT) auftreten, ähnlich wie sie etwa bei PASSAT zu beobachten sind (allerdings ist die Strategie bei SATAN etwas komplexer, so dass weniger Zerlegungsfehler auftreten, vgl. auch Anm. 4.14);

- eine Derivationsanalyse (derzeit nur rudimentär entwickelt als Suffixanalyse);

- die Verarbeitung 'unbekannter' Wortformen (partiell werden dabei Endgrapheme zur Reduktion von potentiellen Mehrdeutigkeiten ausgenutzt).

Diese Vorgehensweise wird auch bei nichtlexikalisierten Einträgen im allgemeinen zu korrekten syntaktischen Strukturbeschreibungen führen; es bleiben jedoch - besonders bei 'unbekannten' Wörtern - Probleme bei der Grundformermittlung und natürlich die semantische Analyse.

Das Ergebnis der semantischen Analyse ist besonders stark abhängig von diesbezüglichen Informationen und Regeln. Sie werden einem Wort zunächst in einem spezifischen semantischen Lexikon zugeordnet. Zu diesen 'Regeln' gehören etwa Angaben über eine semantische und z.T. auch die syntaktische Verträglichkeit zwischen einem 'Prädikat' (Verb, Adjektiv) und seinen Argumenten (z.B. Nomina, Nebensätze). Dies kann v.a. zur semantischen Disambiguierung von Wörtern benutzt werden. Voraussetzung ist u.a., dass z.B. die Nomina 'semantisch' klassifiziert werden (x ist belebt/unbelebt; abstrakt/konkret ...). Daneben werden mithilfe dieses Lexikons v.a. verbhaltige Feste Wendungen zu identifizieren versucht, indem der lexikalische Kontext und dessen mögliche syntaktische Struktur herangezogen werden (Beispiel.: BEWAHREN -- VERSCHWIEGENHEIT (Akk); SETZEN -- (IN) KRAFT). Hier tritt also wiederum, besonders bei fachgebietsorientierten Angaben - etwa zur Ermittlung von 'Festen Wendungen' -, ein nicht unbeträchtlicher Pflegeaufwand für den potentiellen JUDO-Anwender auf.

Mittelfristig muss daher erreicht werden, dass der Aufbau der Lexika möglichst benutzerfreundlich ist. Ansätze dazu werden in der Laborphase von JUDO erprobt (vgl. Kap. 3.1); hierauf müsste beim Übergang in eine 'Produktionsphase' besonderer Wert
gelegt werden.

1.3.3 Textaufbereitung und Analyse

Es wurde schon darauf hingewiesen, dass sich die Probleme der maschinellen Sprachverarbeitung vereinfachen (oder umgehen lassen), wenn die textuellen Restriktionen verstärkt, d.h. die zu verarbeitenden Texte in Struktur und Inhalt begrenzt werden.

Mithilfe des JUDO-Systems sollte zunächst versucht werden, Texte ohne irgendwelche Restriktionen zu verarbeiten. Dabei ist man - zumindest im Labormodell und im Hinblick auf die Verwendung von SATAN - an einige technische und strukturelle Grenzen gestoßen. Beispielsweise ist die maximale Satzlänge bei SATAN auf 100 Wörter begrenzt; in den zu verarbeitenden Texten (u.a. dem Bundesdatenschutzgesetz, das im Test von JUDO derzeit im Mittelpunkt steht) sind jedoch längere Sätze durchaus - wenn auch seltener - möglich. Ähnliches gilt für überlange Wörter: bei SATAN darf ein Wort im Text nicht mehr als 40 und ein Lemmaname (im Sinne von JUDO ein 'Deskriptor') nicht mehr als 36 Buchstaben lang sein.

Diesen bei entsprechender technischer Organisation (variable Satz- und Wortlänge) durchaus lösbaren Problemen stehen eine Reihe von texttypologischen Schwierigkeiten gegenüber, die nur mit editorischen und strukturverarbeitenden Verfahren zu bewältigen sind und gegenwärtig beträchtliche Restriktionen darstellen.

Beispiele für diese komplexen formalen Textstrukturen, die über 'normale' Satzbauregeln (mit der üblichen, im Deutschen an sich schon komplizierten Zeichensetzung) allein nicht zu bewältigen sind, finden sich in den analysierten Texten (vielleicht überproportional) in großer Zahl (vgl. Abb. 1.2: Originaltext des BDSG §§ 1 und 2). Es handelt sich v.a. um spiegelstrichartige oder numerisch gekennzeichnete Aufzählungen, um Hervorhebungen von Überschriften durch Schriftarten und -größen und um Klammerausdrücke unterschiedlichster Art.

Ein Ausweg - wie er auch im Laborsystem JUDO verwendet wird - ist die intellektuelle Präkodierung von Texten (z.B. explizite Markierung des Satzendezeichens, Kennzeichnung von Einschüben: vgl. Abb. 1.3: Text-Präkodierung). Mittelfristig müssen jedoch dringend algorithmische Verfahren gefunden (und integriert) werden, die den Effekten der Präkodierung gleichkommen. Nach den bisherigen Erfahrungen wird in diesem Bereich von der Anwenderseite auch auf längere Sicht noch ein gewisser Anpassungsaufwand erforderlich sein; umgekehrt werden sich die Sprachanalyseverfahren in Zukunft verstärkt an den durch diese Strukturierung vermittelten Zusatzinformationen ausrichten. Z.B. wird die Information, dass es sich bei dem Textteil um eine Überschrift handelt, besondere Analysestrategien anstoßen. Letztlich gelten diese Aussagen auch für Informationen zur Textsorte: Eine Norm (mit Paragraphengliederung) wird andere Verweisanalysen erfordern als etwa ein Literatur-Abstract.

Mit der starken Betonung der Textaufbereitungs- und -typisierungskomponente soll nicht der Eindruck erweckt werden, als brauche nur hier etwas mehr getan zu werden (sei es intellektuell oder maschinell), um die Probleme der maschinellen Sprachverarbeitung zu bewältigen. Manche Fragen linguistischer Art sind zumindest vor dem Hintergrund der schon fast traditionellen syntakto-semantischen Sprachverarbeitung, zu denen auch das System SATAN zu rechnen ist, derzeit nicht gelöst oder prinzipiell nicht lösbar. Wenn man erfahren hat - und dazu ist das JUDO-Modell ein wichtiges Instrument -, dass Qualität und Effizienz der Sprachanalyse (d.h. das Erkennen einer vorliegenden Struktur und die meist damit verbundene Möglichkeit, eine semantische Disambiguierung vorzunehmen) doch erheblich auch von solchen verhältnismäßig trivialen Problemen abhängig sind, darf man diese Komponente - die für die benutzerseitige Akzeptanz u.U. entscheidend ist - nicht vernachlässigen.

Es soll im Rahmen der Übersicht im folgenden nicht weiter auf die Strategie des linguistischen Analyseverfahrens eingegangen werden, die dem System SATAN (und damit dieser Teilkomponente von JUDO) zugrundeliegt. Dazu sei auf das SATAN-HANDBUCH 1978ff (Loseblattsammlung mit laufender Ergänzung/Änderung) und auf die Kap. 3 und 5 verwiesen.

1.4 Auswertungsmöglichkeiten für die Referenz-Dokumentation

Interessant ist im vorliegenden Zusammenhang die Output-Schnittstelle zur Weiterverarbeitung der Analyseergebnisse im Rahmen von JUDO.

Es wurde bereits darauf hingewiesen, dass bei JUDO derzeit nur ein Teil der Analyseergebnisse von SATAN verwertet wird. Es handelt sich insbesondere um die Auswertung der ermittelten Nominalstrukturen und der syntaktischen bzw. semantischen Disambiguierungen. Da es mit SATAN möglich ist, deutsche Texte mit oder ohne Berücksichtigung der Groß-Klein-Schreibung am Wort-/Satzanfang zu verarbeiten, ist auch die Behandlung von Texten möglich, in denen die Substantive nicht durch Großschreibung im Satzinnern besonders gekennzeichnet sind (übrigens sind hierbei allenfalls längere Analysezeiten, aber kaum größere Fehlanalysen festzustellen).

Die (syntaktische) Strukturanalyse von SATAN liefert u.a. - wie schon erwähnt - Nominalstrukturen. Beispiele aus dem Bundesdatenschutzgesetz (BDSG), das im Mittelpunkt des Labortests steht, sind (vgl. z.B. § 1 BDSG) Nominalstrukturen der folgenden Art:

- AUFGABE DES DATENSCHUTZES (Genitiv-Attribut)
- PERSONENBEZOGENE DATEN (Feste Wendung)
- SPEICHERUNG, ÜBERMITTLUNG (nominale Anreihung)
- AUTOMATISIERTE VERFAHREN (Adjektiv-Attribut)
- SCHUTZ VOR MISSBRAUCH (Präpositional-Attribut)

Derartige Strukturen werden - wie erwähnt - von JUDO als Komplexe Deskriptoren zu einem Dokument vergeben; zugleich werden auch die Einzelworte als (Normal-)Deskriptoren verfügbar gemacht (vgl. die einfachen und Komplexen Deskriptoren in den Abb. 6.1, 6.2 und 6.6 bis 6.14).

Bei Komposita und Festen Wendungen werden (über entsprechende lexikalische Angaben) auch die für sinnvoll erachteten Bestandteile vergeben. Ausgangspunkt für eine Vergabe/Zerlegung ist die Entscheidung der Fachwissenschaftler: bei PERSONENVEREINIGUNG wird also etwa PERSON und VEREINIGUNG vergeben, wobei jedoch markiert wird, dass es sich bei den Simplizia bzw. Teilworten um eine durch Zerlegung ermittelte Form handelt.

Bei semantisch mehrdeutigen Wörtern wird auf zweierlei Weise eine kontextbezogene Disambiguierung versucht (vgl. Kap. 3.3):

Im Wesentlichen werden linguistische Verfahren angewendet. Hier helfen zum Teil syntaktische Regeln (z.B. 'jemand HANDELT' vs. 'es HANDELT sich'), zum Teil der nähere semantische Kontext, unter Umständen auch Einzelwörter, z.B. die ÜBERMITTELNDE STELLE (im Sinne von 'Institution') vs. die VERMITTELTE STELLE (im Sinne von 'Job').

In einer Reihe von Fällen reichen derartige (in SATAN einzubringende) Regeln nicht aus, da der Kontext - zur Zeit vor allem der Satzkontext - zu neutral für eine Disambiguierung ist, allenfalls können einige Bedeutungsvarianten ausgeschieden werden. Geht man von den sprachlich möglichen Bedeutungen von Wörtern aus, so scheint es, dass die Auflösung von Bedeutungsvarianten in vielen Fällen nicht über formallinguistische Regeln zu bewältigen ist.

Untersuchungen zu Fachteilgebieten - im vorliegenden Falle zum Datenschutzrecht - haben jedoch gezeigt, dass ein Großteil dieser potentiellen (d.h. lexikalischen) Bedeutungen eines Wortes in einem Fachgebiet - unter probabilistischen Gesichtspunkten - nicht oder kaum nachweisbar ist, dass vielmehr nur wenige (häufig nur eine, allenfalls zwei bis drei) fachgebietsbezogene Präferenzen gegeben werden können. In allen Fällen, in denen linguistische Verfahren wenig ergiebig oder zu aufwendig erscheinen, kann daher dieses Wahrscheinlichkeitskriterium - wenn auch nicht absolut zuverlässig - zur Disambiguierung herangezogen werden.

Erste Erfahrungen mit diesem Vorgehen (in Ergänzung zu den linguistischen Methoden) zeigen viel versprechende Ergebnisse (vgl. KOPELENT 1979). Gegenwärtig wird eine Angabe zu der Wahrscheinlichkeit zugrundegelegt, mit der eine Bedeutungsvariante in einem Fachtext auftritt. Die Angabe '0%' heißt z.B., dass die Wahrscheinlichkeit des Vorkommens dieser Bedeutungsvariante in dem betreffenden Text äußerst gering ist, so dass diese Bedeutungsvariante bei der Disambiguierung ausgeschlossen werden kann.

Dieses Verfahren ist derzeit noch wenig ausgereift. Wertvollere Hilfen bei der Vereindeutigung von Deskriptoren dürfte ein fachgebietsbezogenes Begriffs-Netzwerk bieten, wenn es auch unter Umständen mit größerem intellektuellen Aufwand für die Erstellung und Pflege eines derartigen Netzes verbunden ist.

Es wäre denkbar (man vgl. in dieser Hinsicht die Feinrecherche von CONDOR: WIELAND 1979), dass zum Vergleich von Suchanfragen und Dokumenten weitere syntaktische Relationen (z.B. Subjekt-Objekt) herangezogen und evtl. auf diese Weise auch - etwa nach dem Muster der CONDOR-Relevanzfunktionen - 'Ähnlichkeiten' zwischen einem Dokument und einer Suchanfrage oder weiteren Dokumenten ermittelt werden. Im Rahmen der zukünftigen Projektentwicklung sind solche Verfahren - soweit sie zur Verbesserung der Retrievalergebnisse beitragen bzw. den Benutzerkomfort erhöhen - ein wesentlicher Untersuchungsgegenstand von JUDO.

In der Erprobung ist im Rahmen von JUDO inzwischen ein Konzept, das es dem Gelegenheitsbenutzer erlaubt, sich aus einer natürlichsprachigen Problembeschreibung automatisch die Deskriptoren ermitteln zu lassen, die den Systemkonventionen des zugrundeliegenden Informationssystems entsprechen. Dabei werden zur Ermittlung die gleichen Verfahrensschritte benutzt, denen auch die Dokumente der Informationsbank zuvor unterworfen wurden.

Die ermittelten komplexen bzw. einfachen (weitgehend auch semantisch vereindeutigten) Deskriptoren sind zur Zeit über ein anschließendes boole'sches Retrieval recherchierbar. Auf dieser Stufe werden bereits einige Vorteile deutlich: Der Benutzer wird automatisch auf korrekte Schreibweisen von Begriffen hingewiesen, mehrwortige Begriffe werden identifiziert und syntaktisch relationierte Begriffe sind entsprechend markiert. Anfrage- und Dokumentenprofil werden dabei weitgehend angenähert (vgl. Kap. 6.2 und die Abb. 6.5 und 6.14).

1.5 Datenbank- und Retrievalsystem

Die durch die Analyse zu einem Dokument (Text) ermittelten Deskriptoren, der zugrundeliegende Dokumenttext und - falls erforderlich - zusätzlich intellektuell vergebene (Struktur-)Deskriptoren werden auf einer weitgehend rechnerunabhängigen (Datei-) Schnittstelle zur Weiterverarbeitung in einem Informations-Retrieval-System bereitgehalten. Realisiert sind derzeit eine TELDOK-Variante und eine GOLEM-Version.

Bei der Konzeption des JUDO-Systems wird davon ausgegangen, dass der Deskriptorenpflege besondere Aufmerksamkeit gewidmet werden muss. Andererseits können nach ersten Erfahrungen (vgl. Kap. 3.4) semantische Relationierungen von Deskriptoren auch zu syntaktischen (!) und semantischen Disambiguierungen beitragen. Insbesondere wird im Rahmen von JUDO auf die fachgebietsbezogene Relationierung von Begriffen besonderer Wert gelegt. Auch hier ist eine von konkreten IR-Systemen unabhängige Datenbasis gewählt worden, um nicht durch Restriktionen des jeweiligen Informations-Systems (etwa im Hinblick auf Deskriptorlänge, Anzahl der Relationen) von vornherein eingeschränkt zu werden.

Folgende semantische Relationen werden derzeit bei JUDO verwendet (vgl. den Auszug aus den Relationendaten in Abb. 5.7 und die ausführliche Beschreibung in Kap. 4.2.4):

- Assoziationsrelation: frei assoziiert
- Synonymierelation: 'streng synonym'
- Synonymie auf orthographischer Ebene: Rechtschreibvarianten, Pluralformen
- Quasisynonymrelation: 'schwach synonym'
- Abkürzungsrelation: Abkürzungsform/Langform
- Teil-Ganzes-Relation
- Ober-Unterbegriff-Relation
- Antonymierelation (Gegenbegriff)
- Fachrelation: juristische Regelung/juristischer Regelungsgegenstand

Daneben werden Derivationsrelationen (z.B. Substantiv-Verb) aufgebaut; spezielle Synonymierelationen sind für Feste Wendungen entwickelt: Genitivattribut-Synonymie, Präpositional-Synonymie, Adjektiv-Nomen-Synonymie und Anreihungs-Synonymie. Außerdem gibt es entsprechende Zerlegungsrelationen, z.B. zwischen einem Kompositum und seinen Bestandteilen.

Zum Abschluss seien Retrieval-Möglichkeiten von JUDO am Beispiel der TELDOK-Informationsbank kurz illustriert:

Eine Recherche kann mit den 'Normaldeskriptoren' erfolgen; Bedeutungsnummern am Wortende kennzeichnen die verschiedenen Bedeutungsvarianten; (mehrwortige) Feste Wendungen werden wie Normaldeskriptoren behandelt.

Mehrdeutige Begriffe in der Suchanfrage ohne Angabe der Bedeutungskennung führen den Benutzer zunächst zu einem 'Informationsdokument', in dem die verschiedenen Bedeutungen (mit ihren Identifikatoren) an Beispielen erklärt sind. Komplexe Deskriptoren (d.h. solche, die auf Grund der ermittelten syntaktischen Relationen gebildet wurden), werden mit ihrem über den Text ermittelten Relator (A, G, K, P) versehen eingebracht; gleiches gilt für die Teilwort-De-

skriptoren aus den zerlegten Komposita und Festen Wendungen. Eine Recherche unter Verwendung der (bei TELDOK nur eingeschränkt möglichen) Thesaurusrelationen (z.B. S für SYN, ABK, DER; F für QUA, ASS) ist möglich.

Wörter, die als mehrdeutig erkannt sind, deren im Kontext aktualisierte Bedeutung jedoch durch die vorliegenden Verfahren nicht eindeutig ermittelt werden konnte, haben den Relator M erhalten. Ein Auszug aus einer Recherche mit JUDOT soll einen kurzen Eindruck von den verbesserten Möglichkeiten beim Retrieval vermitteln (vgl. Abb. 6.6 ff sowie die Kommentierung' in Kap. 6.2).

Bereits die vorliegende Datenmenge (Bundesdatenschutzgesetz, mehrere Landesdatenschutzgesetze), an der gegenwärtig das System JUDO erprobt wird, zeigt auf, welche Möglichkeiten eine maschinelle Sprachanalyse zu eröffnen vermag. Eine Reihe linguistischer Fragen, aber auch technischer Probleme muss jedoch noch bewältigt werden; mit verbesserten Analysesystemen für die Verarbeitung von textuellen Massendaten - wie sie sich bereits bei der Konzeption des multilingualen computergestützten Europäischen Übersetzungssystems EUROTRA abzuzeichnen beginnen - können weitere Schritte auch im Hinblick auf eine verbesserte Dokumenterschließung getan werden.

Im Rahmen von JUDO-DS soll in den nächsten beiden Projektjahren 1980/1981 unter Weiterentwicklung der linguistischen und informationstechnologischen Softwarebasis ein breiterer Test ergeben, unter welchen Voraussetzungen ein computergestütztes Indexierungs- und Retrievalverfahren auf linguistischer Grundlage realisiert werden kann.

1.6 Zusammenfassung der Ergebnisse

Im Folgenden werden die bisher erreichten Ergebnisse zusammenfassend dargestellt. Dabei wird auf die im Projektantrag aufgeführten Teilkomponenten Bezug genommen und kurz auf Alternativen zu linguistischen Verfahren eingegangen; schließlich werden die Veränderungen (Einschränkungen und Erweiterungen) gegenüber dem ursprünglichen Konzept beschrieben.

1.6.1 Nachweis der Einsatzfähigkeit linguistischer Verfahren in der Dokumentation

Unter 'Einsatzfähigkeit' wird i.W. die Operationalität bzw. Funktionalität der konzipierten Verfahren verstanden. Weder ist dabei an den entsprechenden Nachweis durch den Einsatz bei einem Anwender noch an einen Recall-Precision-Test gedacht.

(1) Maschinelle Zuordnung von Flexionsformen zu ihrer Grundform (Lemmatisierung)

Diese Komponente war zu Projektbeginn anderweitig nur in Ansätzen entwickelt und eher ad hoc behandelt worden (vgl. den Schott-Algorithmus, GOLEM/PASSAT, aber auch STAIRS-TLS). Bei JUDO/SATAN erfolgt die Grundformenermittlung auf systematisch-linguistischer Grundlage. Voraussetzung ist (inzwischen in JUDO implementiert) eine syntaktisch-morphologische Kodierung eines einzigen Stammeintrags je Wort (z.B. auch bei unregelmäßigen Verben). Dieses Verfahren ist inzwischen einsatzfähig (vgl. Kap. 3.1).

(2) Syntakto-semantische Disambiguierung

Das entsprechende Verfahren ist (partiell mittels SATAN, partiell durch computergestützte Nachbereitung im Rahmen von JUDO) implementiert; für den Bereich Datenschutz sind entsprechende Lexika und Regelsysteme im Rahmen von JUDO entwickelt und im Einsatz. Funktionell ergeben sich noch einige Detailprobleme, z.B. bei der Disambiguierung von Adjektiven. Die Erkennung diskontinuierlicher Elemente (z.B. Verbzusatz) ist ebenfalls implementiert (zur Qualität vgl. Kap. 3.2 und Kap. 3.3).

(3) Fachsprachspezifische Wörterbuchmarkierung

In das Gesamtsystem sind - damit über SATAN hinausgehend - fachsprachspezifische Wörterbücher und Relationen integriert, deren Auswertung eine Steuerung der Indexierung, z.B. die Einbeziehung von Kompositazerlegungen in den Disambiguie-rungsprozess und bei der Deskriptorvergabe oder die Verwertung von Wahrscheinlichkeitsgewichtungen bei der Disambiguierung, zulässt. Die Möglichkeit der Eliminierung nicht-sinntragender Wörter bei der Vergabe von Deskriptoren ist gegeben (vgl. Kap. 4.2.3 und Kap. 5.3).

(4) Integration der Analyseergebnisse in ein maschinelles Indexierungssystem

Anstelle des noch im Projektantrag vorgesehenen Systems LEDOC, einem inzwischen am Rechenzentrum der Universität Regensburg nicht mehr gepflegten Informations-Retrieval-System mit Katalog-, aber ohne Dialogfunktionen wurde im Projektverlauf (aufgrund des installierten Rechners Telefunken TR 440) das IR-System TELDOK implementiert (vgl. Kap. 6.2). Die Implementierung einer GOLEM-Datenbank ist inzwischen ebenfalls - mit den vorgesehenen Funktionen - in erster Erprobung (vgl. Kap. 6.1).

Die im Antrag angesprochenen kompatiblen Datenschnittstellen, z.B. zwischen SATAN und JUDO, sind in Form von Deskriptoren-Basissätzen und Relationssätzen eingerichtet (vgl. Kap. 5.2).

(5) Anschluss an ein Konkordanz-System

Anstelle des im Projektantrag vorgesehenen COBAPH-Anschlusses wurde ein eigener Operator geschaffen, der die wesentlichen COBAPH-Funktionen, z.B. den Aufbau einer Konkordanz auf Mikrofiche, enthält. Damit sind auch Off-line-Funktionen von JUDO realisiert, die eine Benutzung der Daten etwa zu Test- und Auswertungszwecken ermöglichen (vgl. Kap. 6.3).

(6) Textbezogene Studien; Vorschläge für die Weiterentwicklung von SATAN

Kontinuierlich wurden entsprechende Studien (SCHNEIDER 1976, J. KOPELENT 1978, W. KOPELENT 1978, HOFMANN 1977) erstellt; die Ergebnisse sind z.T. bereits in SATAN/JUDO implementiert. Vorschläge für die Weiterentwicklung von SATAN und Korrekturen/Ergänzungen (z.B. Verarbeitung von Klammerausdrücken und Einbettungen, Festen Wendungen, angereihten mehrwortigen Begriffen, Integration von Thesaurusrelationen in die Analyseverfahren, Behandlung von Partizipialkonstruktionen, juristischen Verweisangaben) sind in stetem Kontakt mit Saarbrücken erfolgt und z.T. bereits integriert.

Erste Erfahrungen zur Brauchbarkeit und zu Problemen bei der Verwendung eines allgemeinen Sprachanalyseverfahrens (wie SATAN) zur Fachtextanalyse sind z.T. bereits publiziert bzw. werden in diesem Abschlussbericht behandelt (vgl. z.B. Kap. 3.2).

(7) Implementierung des Gesamtsystems

Die Implementierung ist erfolgt und funktional an Dokumenten des Informationsrechts ausgetestet. Insgesamt ist damit der Nachweis bezüglich einer funktionalen Einsatzfähigkeit linguistischer Verfahren erbracht. Der qualitative Nachweis soll bei JUDO-DS (1980/1981) im Rahmen einer Evaluierung geführt werden.

1.6.2 Alternativen zu linguistischen Verfahren

Das Ziel des JUDO-Projekts nachzuweisen, dass linguistische Verfahren gegenüber anderen - z.B. rein statistischen - Verfahren Vorzüge besitzen, ist einzuschränken auf Teilbereiche wie Wortformenreduktion, automatische Auflösung von Mehrdeutigkeiten und fachtextbezogene Kompositazerlegung.

So wurde z.B. die computergestützte Thesauruserstellung in diesem Zusammenhang zunächst (im Projektantrag) nicht angesprochen, auch nicht die Frage der automatischen Deskriptorgewinnung i.S. einer Gewichtung: in diesen Bereichen sind statistische Methoden (als Ergänzung bzw. Grundlage eines Indexierungssystems) äußerst sinnvoll und nützlich. Als Vergleichsbasis war (implizit) an Verfahren wie CONDOR (mit Einschränkung: Lemmatisierung/Derivation), TUBIBMUE, das ehem. Bonner Morphologieprojekt am IKP und PASSAT (im Hinblick auf Zerlegungen) gedacht.

Im Rahmen von JUDO sollten dabei durch die kontrollierte Erzeugung der gewünschten Lemmaform bzw. des gewünschten Zerlegungsbegriffs benutzer- und anwendungsorientierte Begriffsformen zugelassen werden. Dies soll nicht heißen, dass andere Verfahren dadurch entscheidend beeinträchtigt werden. Ein erster Vergleich (anhand des BDSG) der 'Deskriptorenlisten' von PASSAT mit den beim JUDO-System erzeugten Stammformen zeigt, dass hier das JUDO-System einige Verbesserungen bringt (vgl. Kap. 7.2). Wie erwähnt, sollen dadurch Alternativ-Verfahren wie CONDOR oder auch PASSAT und STAIRS/TLS keineswegs im Hinblick auf eine sinnvolle Verwendbarkeit in Frage gestellt werden: Es muss deutlich gemacht werden, dass mit JUDO nur ein bestimmter Teilbereich der Dokumentation erschlossen werden soll (auf Kosten eines intellektuellen und auch softwaretechnologischen Mehraufwands, vgl. jedoch Abb. 4.4), allerdings ein Bereich, der besonders benutzernah und benutzerfreundlich erscheint (vgl. die Diskussion der Reduktion des intellektuellen Aufwands auf Benutzerseite in Kap. 4.1.1). Ein absoluter Anspruch - etwa bezogen auf die Verfahrensweise - kann schon deswegen nicht sinnvoll sein, weil von Anfang an bekannt war oder doch als bekannt vorausgesetzt werden konnte, dass vom späteren Anwender ein gewisser zusätzlicher Arbeitsaufwand - etwa im Bereich der Lexikonerstellung - wohl auf längere Sicht erwartet werden muss. Dennoch wird auch in Zukunft bei JUDO in erster Linie Bezug genommen auf die Qualität der Ergebnisse und nicht notwendig auf den erforderlichen Aufwand.

Nicht in den Projektantrag von 1977 eingebracht war eine ausführliche Bewertung - etwa über Recall-Precision-Untersuchungen (i.S. der Definition in Kap. 1.2.1). Dies konnte allein schon

aufgrund der geringen Zahl von Projektmitarbeitern und wegen des ebenfalls geringen Umfangs der im Projektverlauf beabsichtigten Datenbasis geschlossen werden. Allerdings reicht auch ein Recall-Precision-Test im angegebenen Sinn allein für die Bewertung der Qualität des Verfahrens nicht aus.

1.6.3 Geänderte Projektziele

(1) Umsetzung auf einen Fremdrechner

Zum Zeitpunkt der Antragstellung stand nicht fest, welcher Rechner in Regensburg zur Verfügung stehen würde. Da es sich schließlich um den gleichen Rechnertyp wie in Saarbrücken handelte (TR 440), entfiel notwendig die (1:1-)Umsetzung (ursprünglich war für Regensburg eine CDC Cyber vorgesehen).

(2) Teildatenmenge zum Steuerrecht

Zunächst war an die Verarbeitung von zwei Fachgebieten gedacht worden. Im Bereich Steuerrecht, der neben dem Datenschutzrecht vorgesehen war, sind die lexikalischen Vorarbeiten zum Syntax-Lexikon (d.h. Analyse bis Operator VERA/ KOMA) abgeschlossen. Die vorgesehene Teildatenmenge ist maschinell lesbar erfasst. Nicht durchgeführt ist der Aufbau eines Steuerrechts-Thesaurus (der allerdings auch nicht Gegenstand des Projektantrags war, vgl. aber Kap. 1.6.4) und die Einbringung der Steuerrechtstexte in ein (weiteres) IR-System.

(3) Fragen der Interaktion Mensch-Computer während des Analyseprozesses

Diese (im Antrag optionale) Komponente wurde im Berichtszeitraum (1977-1979) zurückgestellt. Allerdings wird gegenwärtig überlegt, zumindest ein Verfahren zur computergestützten Restriktion, zur Selektion oder Auflösung maschinell nicht vereindeutigter Begriffe/Deskriptoren zu entwickeln.

(4) Exemplarische statistische Angaben

Zu Rechenzeitverhalten, zu Datenumfang, zu Auswirkungen (auch Fehlern) deskriptor-relevanter Homographieauflösung, zu deskriptor-relevanten Nominalgruppen, zu semantischer Vereindeutigung auf SEDAM-Ebene und zum CUT-OFF-Verfahren konnten erst vorläufige Statistiken erstellt werden. Diese Daten wurden zudem aus einer Teilmenge des Textmaterials gewonnen.

1.6.4 Ergänzende Arbeiten

(1) Thesaurus-Relationierung

Aufbauend auf den Richtlinien von DIN 11463 und darüber hinausgehend wurde für die Fachbegriffe des Informationsrechts ein Thesaurus entwickelt und in die aufgebauten IR-Systeme integriert (vgl. die Kap. 14.2.4, 5 und 6).

(2) Aufbau komplexer Deskriptoren

Aufbauend auf den Ergebnissen der linguistischen Analyse wurde ein System sog. 'Komplexer Deskriptoren' entwickelt, das präzisere Dokumenterschließungen und einfachere bzw. genauere Retrievaltechniken erlauben soll (vgl. die Kap. 3, 5 und 6).

(3) Natürlichsprachige Problembeschreibung

Es ist möglich, Deskriptorenlisten (einfache und Komplexe Deskriptoren in der Form, wie sie die IR-Systeme als Suchbegriffe akzeptieren) aufgrund natürlichsprachiger Problembeschreibungen zu erzeugen. Dieses Verfahren ist zur Zeit für die TELDOK-Variante implementiert (vgl. Kap. 6.2.3).