In: Kuhlen, R. (ed., 1979): Datenbasen, Datenbanken, Netzwerke. Praxis des IR. Band 1, Aufbau von Datenbasen. München et al: KG Saur, 311-338

Harald Zimmermann

Ansätze einer realistischen automatischen Indexierung unter Verwendung linguistischer Verfahren

1. Zum Begriff der maschinellen Indexierung
2. Anwendungsbereiche und allgemeine Problemstellung
    2.1 Freitextverarbeitung
    2.2 Klartextverarbeitung
3. Probleme der maschinellen Sprachanalyse
4. Anforderungen an IR-Verfahren mit linguistischer Komponente
5. Vergleich von Verfahren und Verfahrensmoduln
    5.1 PASSAT
    5.2 CONDOR
    5.3 JUDO

Zusammenfassung
Der IST-Zustand der automatischen Indexierung im Rahmen von Referenz-Retrieval-Systemen (PASSAT) wird mit den Perspektiven eines Systems der achtziger Jahre (CONDOR) kontrastiert. Die Möglichkeiten linguistischer Methoden werden erörtert. Im Zentrum stehen Ausführungen zu JUDO, das unter Ausnutzen und Weiterentwickeln der in Saarbrücken erarbeiteten Sprachanalyseverfahren juristische Fachtexte automatisch indexieren soll. Es wird ein streng 'lexikalischer' Ansatz verfolgt. Syntaktische und semantische Disambiguierung wird besonders dargestellt. Verschiedene Klassen von im Retrieval möglichen Deskriptoren werden eingeführt.

Abstract
The current state of automatic indexing within the framework of reference retrieval systems (PASSAT) is contrasted with the perspectives afforded by a system for the 80's (CONDOR). The possibility of using linguistic methods is discussed. Of central importance are comments an JUDO, which, as a further development of the language analysis procedures worked out in Saarbrücken, is intended to index legal texts automatically. A strictly 'lexical' approach is used. Syntactic and semantic disambiguation are accorded special attention. Different classes of possible descriptors in retrieval are introduced.

1. Zum Begriff der maschinellen Indexierung

Unter Indexierung wird streng genommen nur die Abbildung eines Dokumentinhalts auf Schlüsselwörter (Stichwörter, Schlagwörter, Deskriptoren oder Klassifikatoren) verstanden. Dabei können diese Schlüsselwörter nach dem Prinzip des coordinate indexing isoliert stehen oder (in verfeinerter Form) zueinander in syntaktische Relation gesetzt werden. Diese Schlüsselwörter dienen beim Retrieval zur Identifikation eines (relevanten) Dokuments. Enthalten die Dokumente natürlichsprachliche Daten (Text), so lassen sich auf diesen Text Verfahren anwenden, wie sie in der maschinellen Sprachverarbeitung entwickelt werden. Wird beim Retrieval (der Dokumentensu-

che) eine natürlichsprachige Problembeschreibung ("Suchfrage") verwendet, sind jedoch prinzipiell die gleichen Methoden anwendbar. Insofern sind einige der nachfolgenden Aussagen nicht auf den eigentlichen Indexierungsteil des maschinellen Prozesses, die Dokumenterschließung, beschränkt.

Im Folgenden wird Indexierung (wie allgemein üblich) im wesentlichen auf referenzielle Abbildungen von Textdokumenten auf Begriffe (Deskriptoren i.w.S.) und das entsprechende Referenz-Retrieval bezogen; als 'Antwort' auf eine 'Suchfrage' wird auf die Adresse oder Kennung von Dokumenten verwiesen, in denen die Fragestellung möglicherweise behandelt ist. Es wird abgesehen von der Faktenerschließung und dem Fakten-Retrieval i.S. konkreter Antworten auf Entscheidungsfragen ('ja/ nein') oder Ergänzungsfragen (z.B. 'wer, wie, wo'). Die Grenzen zwischen Indexierung und Faktenerschließung im Bereich der linguistischen Dokumentverarbeitung, wenn man überhaupt von solchen sprechen kann, können allerdings nicht so klar gezogen werden.

So wünschenswert eine maschinelle Erschließung von Fakten eines (Sprach-)Dokuments sein mag, so interessant diese Aufgabe wissenschaftlich ist: es kann gleich festgehalten werden, dass für die derzeitige Praxis des Information Retrieval derartige Verfahren (noch) nicht relevant erscheinen: Einerseits liegt dies an den (mitunter noch) unzureichenden sprachwissenschaftlichen Erkenntnissen, aber auch an dem erheblichen Kostenfaktor einer entsprechenden Dokumentaufbereitung. Der intellektuelle Aufwand, der für eine (dann meist noch immer mangelhafte) maschinelle Fakten-Erschließung betrieben werden müsste, liegt sicherlich weit über dem für eine entsprechende und qualitativ höhere rein intellektuelle Erschließung. Allerdings kann man feststellen, dass Verfahren, die sich - grob gerechnet - dem Bereich von natürlichsprachlichen Projekten der sog. 'künstlichen Intelligenz' zuordnen lassen, gegenwärtig (d.h. zum Ende der 70-er Jahre) begonnen haben, über das reine Interesse an (sprach-)wissenschaftlichen Erkenntnissen hinaus auch für die Anwendung im Information Retrieval Wirkung zu zeigen; allerdings bedürfen derartige Labormodelle (vgl. z.B. das Verfahren PLIDIS des IDS, Berry-Rogghe/Wulz 1976) noch intensiver - auch grundsätzlicher - Diskussion und Erprobung.

2. Anwendungsbereiche und allgemeine Problemstellung

Für die vorliegende Studie stelle man sich - etwas vergröbernd - folgende Situation vor: Ein textuelles Dokument - darunter kann ein Brief, ein Protokoll, ein Bericht, eine Richtlinie, ein Abstract, ein Befund, ein Gesetzestext (Paragraph), eine Notiz, eine Softwarebeschreibung usf. verstanden werden - soll einem computergestützten IR-Prozess unterworfen werden.

Für den Fall des Einsatzes eines automatisierten linguistischen Verfahrens ist der Text des Dokuments maschinenzugänglich (d.h. maschinenlesbar und maschinenverarbeitbar) zu erfassen. In manchen Fällen, vor allem dann, wenn eine entsprechende Technologie zur Verfügung steht (Lichtsatz, Text- oder Erfassungssystem, ggf. auch eine für optisches Lesen geeignete Schreibmaschine-Konfiguration) stellt das ersterzeugte Produkt bereits eine mögliche Basis zur maschinellen Weiterverarbeitung dar. Doch selbst unter günstigen technologischen Voraussetzungen ist organisatorisch nicht immer die Möglichkeit gegeben, diese Dokumente von der äußeren Form her (Aufbau, Gliederung) zu beeinflussen. Dies hat m.E. gegenwärtig eine größere Bedeutung für die praxisorientierte Sprachverarbeitung, als allgemein angenommen wird. Herrscht schon im Hinblick auf den Dokumenttyp und die formale Gestaltung eines Textes eine große Vielfalt, so

gilt dies auch für Faktoren wie den möglichen Themen- und Fachbereich, die Ausdrucksweise ("Stil") und die Korrektheit der Daten (Fehler im Bereich der Rechtschreibung, der Zeichensetzung, der innersprachlichen 'Logik'). Dies mag zur Illustration genügen.

Zu Beginn der Überlegungen beim Aufbau eines textuellen IR-Systems sollte daher das Textmaterial aufgrund derartiger Kriterien entsprechend aufbereitet werden. Diese Textaufbereitung ist eine wesentliche Voraussetzung für den Erfolg des ggf. anzuwendenden Verfahrens, zumindest haben solche Maßnahmen (z.B. vorheriges Korrekturlesen, weitere Präkodierung der Texte) auf die Ergebnisse der computergestützten Indexierung wesentlichen Einfluss. Derartige Fragen, etwa auch das Problem einer dokumentationsgerechten 'Paraphrasierung' (naturgemäß mit einigen Konsequenzen, z.B. Verlust der 'Orginalität') sind m.W. noch nicht intensiv untersucht worden. Allenfalls hat man sich mit dem Problem der 'automatischen' Rechtschreibfehlererkennung (oder auch -korrektur) befasst (vgl. z.B. Wingert 1978).

2.1 Freitextverarbeitung

Einerseits muß davon ausgegangen werden, dass der Administrator eines IRS (z.B. der für den Aufbau eines IRS zuständige Dokumentar) keine Einwirkungsmöglichkeiten auf die Gestaltung des Original-Inputs hat bzw. keine solchen Einflüsse - abgesehen von Sachzwängen, die sich aus einem Thema ergeben - vorgegeben sind. Dies bedeutet, dass sich der Autor einer völlig freien Ausdrucksweise hat bedienen können, dass der verwendete Wortschatz nicht explizit kontrolliert und auch die äußere Form des Dokuments nicht vorgeschrieben ist.

Eine maschinelle Aufbereitung derartiger Dokumente wird i.a. mit "Freitextverarbeitung" bezeichnet (1); Beschränkungen dabei können allenfalls über "natürliche" Prozesse vorliegen (Beispiele dafür sind etwa ein spezifisches Fachgebiet, eine bereits stark konventionalisierte (oft verknappte) Darstellung, z.B. Wetterbericht, aber auch Dokumente wie Abstract, Diagnose, Befund; vielleicht auch einige Fachtexte). Im letzten Fall ist allerdings kaum geklärt, inwieweit echte Einschränkungen vorliegen, eher kommen Komponenten 'erschwerend' hinzu wie Formelsprache, Tabellen, Verweissprache (im Recht z.B.).

2.2 Klartextverarbeitung

Einen gewissen Kontrast dazu bildet die sog. "Klartextverarbeitung" (2). Hierbei unterwirft sich ein Autor bewusst bestimmten Einschränkungen und "künstlichen" Konventionen, z.B. im Hinblick auf den zu verwendenden Wortschatz oder die Vermeidung komplexerer syntaktischer Strukturen; u.U. muss er dabei auch "äußerlichen" Konventionen (z.B. Satzlängenbeschränkung) folgen. Derartige Regulierungen müssen nicht notwendig für einen menschlichen Kommunikationspartner sichtbar werden, sie tun es aber häufig. Diese Methode ist übrigens nicht auf die linguistische Datenverarbeitung beschränkt; ihre Ursprünge finden sich vielmehr im Bereich der Humanübersetzung und des Fremdsprachenunterrichts (z.B. Basic English) (3).

Es wurde schon betont, dass die Probleme der maschinellen Sprachbearbeitung in dem Maße sich vereinfachen, in dem man in geeigneter Weise die textuellen Restriktionen verstärkt.
Im Extremfall kann man manche Probleme der maschinellen Sprachverarbeitung sogar völlig 'überspielen'. Ein charakteristisches Beispiel dafür ist TITUS-2 (Zingel 1978). Durch Kontrolle

des 'Vokabulars' werden Bedeutungs-Mehrdeutigkeiten ausgeschlossen, durch die strenge syntaktische Formatierung (Vorgabe von Satzbauplänen mit Valenzkennungen) und durch explizite Präkodierung morphologischer Informationen wird ereicht, dass mehrsprachige 'Übersetzungen' (man würde besser von 'Umsetzungen' reden) - übrigens auch in die jeweilige Textoberfläche der Ausgangssprache - erfolgen. Hier genügte ein der traditionellen Linguistik entlehntes Konzept - man kann etwa Teile der Dependenzgrammatik (v.a. den Valenzrahmen) als 'Vorbild' erkennen -, um das Problem der Mehrsprachigkeit von Abstracts zu bewältigen, wenn auch unter größerer Belastung des intellektuellen Indexierers (4). Erstaunlicherweise sind die Strukturierungen von TITUS-Daten bislang nicht im eigentlichen IR-Prozess verwendet worden, obwohl sich dies anbietet, insbesondere im Hinblick auf die Verwertung (explizit kodierter) syntakto-semantischer Relationen.

3. Probleme der maschinellen Sprachanalyse

Wenn man "ernsthafte" linguistische Verfahren bei der computergestützten Indexierung heranziehen will, steht man vor einer Reihe von Problemen, deren Lösung i.w. von der Bewältigung zweier Faktoren abhängig ist:

> (i) die "natürliche" Mehrdeutigkeit syntaktischer und semantischer (Teil-)Strukturen (im einfachsten Fall sind es Wörter); sie ist letztlich nur über eine 'tiefe', d.h. eingehende sprachliche Analyse der Dokumente zuverlässig möglich; rein (traditionell) linguistische Beschreibungen reichen dazu sicher nicht aus, hier wird letztlich die Einbeziehung von (enzyklopädischem) Weltwissen, von psychischen oder sozialen Faktoren u.s.f. nötig sein.
>
> (ii) Selbst wenn es gelingt, die jeweilige Sprache (über ein komplexes formales System) so weit zu beschreiben, dass eine Auflösung aller oberflächenstrukturell vorgegebenen Mehrdeutigkeiten von Texten/Sätzen maschinell möglich erscheint, bleibt die Frage der Strategie zu klären, nach der diese Analyse/Beschreibung erfolgen kann. Die Komplexität struktureller Möglichkeiten, bedingt z. B. durch Oberflächentilgungen oder durch (häufige) partielle syntaktische Mehrdeutigkeiten, ist mit heutigen Mitteln der Datenverarbeitung in vertretbarem Zeitaufwand kaum zu bewältigen - immer unter der Voraussetzung, dass man so lange 'analysiert', bis man sicher ist, jede noch mögliche Mehrdeutigkeit aufgelöst bzw. jede relevante aktuelle Struktur (d.h. die adäquate Beschreibung) ermittelt zu haben.

Die Lösung eines wesentlichen Problems der Indexierung, die Themenschwerpunkte eines Dokuments zu erschließen, d.h. eine Konzentration der Information (Verdichtung, Befreiung von Ballast) zu erreichen, wird auf 'linguistischem Weg' (besser: ohne statistische Verfahren) erst möglich sein, wenn eine verstehensadäquate maschinelle Sprachverarbeitung möglich ist.

Ähnlich problematisch erscheint derzeit die automatische Ermittlung semantischer Relationen (Synonymie, Ober-Unterbegriff, prinzipiell ebenfalls ein wichtiges Desiderat der Indexierung) aufgrund linguistischer (besser wiederum: nichtstatistischer) Verfahren.

Die bisher bekannt gewordenen Verfahren erheben alle nicht ernsthaft den Anspruch, eine absolute 'Verstehensadäquatheit' zu erreichen. V.a. diejenigen Analysesysteme, die eine sehr feine und tiefgehende Textbeschreibung zum Ziel haben, sind jedoch aufgrund der ansonsten gegebe-

nen Komplexitäts-Explosion zu Strategien gezwungen, die die adäquat zu verarbeitenden Texte erheblich restringieren. Im allgemeinen werden dabei Texte in zweierlei Hinsicht restringiert:

(i) Beschränkung des Themen-/Sachbereichs (Aufbau einer Mini-Welt, z.B. - vereinfacht dargestellt - einem Zimmer mit zwei Stühlen, einer Spielzeugbox mit kombinierbaren Bauklötzchen, einer Aktienbörse (steigende / fallende Aktien), einer Reisesituation (z.B. Name, Ort des Reisenden usw.)

(ii) Beschränkung der sprachlichen Möglichkeiten (z.B. eingeschränkte Syntax, kaum semantische Mehrdeutigkeiten).

Derartige Verfahren der 'Künstlichen-Intelligenz-Forschung' sind ohne Zweifel von hohem Erkenntniswert, sie simulieren aber nur wenige und sehr einfache reale Situationen, und wo sie interessante Themen ansprechen, sind meist weitaus effizientere Verfahren in der Praxis denkbar (z.B. graphischer gegenüber natürlichsprachlichem Dialog, feste Anzahl von Suchfragen; kompakte formale Darstellung). Eine Reduktion der vorgegebenen Restriktionen kann kaum erfolgen; die meisten Anwendungssituationen, bei denen sich der maschinelle Einsatz kostenmäßig vertreten ließe, sind wegen ihrer größeren Komplexität nicht zu adaptieren.

Dies soll nicht heißen, dass sich bei intensiver Suche nach 'Freiräumen' für derartige Verfahren nicht auch geeignete IR-Objekte finden ließen. Man ist in der typischen IR-Situation jedoch meist gezwungen, große Datenmengen zu verarbeiten, der Wortschatz ist allenfalls geringfügig einschränkbar, die Verarbeitungszeiten dürfen nicht wesentlich über den bisher bekannten Werten liegen (etwa verglichen mit STAIRS oder GOLEM/PASSAT); der Aufwand für die Implementierung oder Wartung des Systems im Bereich der maschinellen Sprachanalyse darf nicht sehr hoch sein, da die "intellektuellen" Strategien (beim Retrieval) kostengünstiger erscheinen, z.B. das Ausfiltern von "Ballast" nach einer Grobrecherche (über Schlüsselwörter, evtl. - wie vielfach üblich - unter Einschränkung der Suchanfrage mithilfe formatgebundener Deskriptoren wie Datum, Sachbereich). Das Bestreben, die Dokumentmenge bei der Grobrecherche bereits auf eine überschaubare Größenordnung zu bringen, kann natürlich auch zum Ausfiltern relevanter Dokumente führen; im (Literatur-)Dokumentationsbereich mag diesem Problem (von der Benutzerseite her) kein allzu großes Gewicht beigemessen werden: allerdings, vergleiche man die Patent- oder Rechtsdokumentation, wo das Bestreben, einen hohen Recall zu erhalten, sehr zu Lasten der Precision geht, also das Erreichen einer umfassenden Information mit dem rapiden Anwachsen von Ballast einhergeht.

Vor dem Hintergrund dieser Aussagen (oder zumindest einer entsprechenden Argumentation der 'Praktiker') können die folgenden Ausführungen betrachtet werden; zumindest wird sich für den Anwender der Begriff einer "realistischen linguistischen Indexierung" auch unter diesem Aspekt manifestieren. Dennoch sollen - nicht ohne auf die angeführte Interpretation hingewiesen zu haben - diejenigen Aspekte im Vordergrund der Betrachtung stehen, die etwa unter dem Begriff der "derzeitigen Leistungsfähigkeit" von linguistischen Methoden im Rahmen von IR-Systemen gefasst werden können. Dafür lassen sich auch ernstzunehmende Gründe anführen: Mit der raschen Entwicklung der Mikroprozessortechnik, der anstehenden Implementation von Großspeichern und den damit verbundenen Kostensenkungen bei gleichzeitiger Steigerung der Rechenleistung wird mittel- bis kurzfristig die Kosten/Nutzen-Frage auch für linguistische Verfahren in einem neuen Licht gesehen werden (5).

4. Anforderungen an IR-Verfahren mit linguistischer Komponente

Man kann davon ausgehen, dass die verschiedenen IR-Systeme in Zukunft um die Gunst des Anwenders verstärkt dadurch wetteifern werden, dass sie den Komfort und die Akzeptanz ihres Systems durch Verbesserung der Benutzerschnittstelle, insbesondere auch im textuellen Bereich, erhöhen werden. Oberflächige (Hilfs-)Methoden wie z.B. die Maskierung von Zeichen werden eine untergeordnete Rolle spielen, wenn z.B. exakte Reduktionsalgorithmen, (z.B. Pluralumlaute: Häuser/ Haus), Kompositazerlegung (Buchprojekt / Buch & Projekt) als Funktionen realisiert sind.

Insgesamt haben die linguistischen Verfahren etwa folgenden Bedingungen zu genügen:

- Verbesserung bei der Präzisierung von Suchfragen (evtl. auch Abkürzung des Suchvorgangs)
- keine Erniedrigung des Recalls (nach Möglichkeit sogar merkliche Erhöhung)
- Vereinfachung der Problembeschreibung beim Retrieval (Reduktion des Formalismus) (6).
- Homogenisierung des Verhältnisses zwischen Dokumenterschließung ('Indexierung') und Dokumentidentifikation ('Retrieval')
- einfache Administration des Systems (der Endbenutzer sollte bei der Pflege bzw. Anpassung des Verfahrens auf "seine" Bedürfnisse keine größeren Kenntnisse, insbesondere keine tieferen Einsichten in die Struktur von Sprache, einbringen müssen).

5. Vergleich von Verfahren und Verfahrensmoduln

Die im folgenden vorzustellenden Verfahren erfüllen mehr oder minder diese Kriterien. Ich beschränke mich dabei auf Entwicklungen und Forschungsprojekte in der Bundesrepublik Deutschland (7). Dabei werden i.w. drei Verfahren vorgestellt und in ihrem Funktionsspektrum verglichen. Es handelt sich einmal um das in der Praxis mehrfach bereits eingesetzte System PASSAT (im Rahmen von GOLEM): es dient v.a. zur Kontrastierung des jetzigen realisierten IST-Standes mit dem zweiten hier vorzustellenden System CONDOR, dessen linguistische Komponente behandelt wird. Für den Anfang der 80-er Jahre ist bei CONDOR eine Produktversion vorgesehen (8), z.Zt. liegt eine Laborversion vor. Im Mittelpunkt steht jedoch die Darstellung der wesentlichen Komponenten des in Regensburg im Aufbau befindlichen Experimentalsystems JUDO (9). Im Rahmen von JUDO wird ein automatisches Sprachanalysesystem verwendet, das an der Universität des Saarlandes entwickelt wird (10). Gemeinsam ist diesen Verfahren, dass sie prinzipiell keine textuellen Restriktionen kennen, d.h. auf beliebige (deutschsprachige) Textdaten angewendet werden können.

PASSAT unterstützt i.w. nur die Ermittlung von Schlüsselwörtern (11). Eine Wortform wird anhand einer Vergleichswortliste (VWL) identifiziert, die bei den hinzukommenden (d.h. in Dokumenten neu auftretenden) Wörtern intellektuell ergänzt werden muss. Wortformen (Flexionsformen) werden dabei auf ihre Grundformen reduziert. Wortzusammensetzungen (Komposita), deren Bestandteile in der VWL enthalten sind, werden als solche ermittelt (dies ist im

Deutschen ein beträchtlicher Anteil). Unflektierte mehrwortige Ausdrücke, die kontinuierlich (d.h. unmittelbar nacheinander) im Text auftreten, können zu einem 'Wort' zusammengefasst werden. Eine (meist intellektuell gepflegte) Assoziationsmatrix erlaubt i.W. die Eliminierung von nicht für dokumentationsrelevant gehaltenen Einträgen (Stoppwörtern).

Die Praxis (z.B. die Anwendung bei JURIS) zeigt, dass die Pflege der VWL doch sehr aufwendig ist. Daneben lassen sich falsche (d.h. unsinnige) Wortzerlegungen nicht ausschließen (z.B. PROZESSHANDLUNGEN in PROZESS, HANDLUNG, HAND, LUNGE); das Problem der Homographie und Homonymie (der Bedeutungsdifferenzierung) ist (natürlich) für den Einzelfall, d.h. die jeweilige Belegstelle nicht gelöst, diskontinuierliche sprachliche Einheiten (z.B. GING ...WEG in WEGGEHEN) werden nicht zusammengefasst u.a.m. Obwohl derartige Mehrdeutigkeiten allenfalls ca. 15% der Substantive und 20% der Verben betreffen, sie darüber hinaus im aktuellen Retrieval durch komplexere Suchanfragen weitgehend eliminiert werden können, ergibt sich eine Reihe von Nachteilen. Einerseits wird die Akzeptanz des Systems beeinträchtigt (12), die in GOLEM mögliche Thesaurus-Relationierung wird behindert, eine Zerlegung von Komposita in "sinnvolle" Bestandteile ist erschwert: (steht z.B. ESEL und BRÜCKE in der VWL, aber nicht ESELSBRÜCKE, so erfolgt "automatisch" eine Zerlegung; sie kann nur durch intellektuelle Kontrolle der Ergebnisse, im Fehlerfall durch Aufnahme des Gesamteintrages (hier ESELSBRÜCKE) in die VWL (u.U. nebst erneuter Dokumentanalyse) vermieden werden.

5.2 CONDOR

Aufbauend auf den Erfahrungen bezüglich der Probleme, die bei Verfahren wie PASSAT auftreten, wurde im linguistischen Teil von CONDOR zunächst ein völliger Neuansatz erprobt. Insbesondere sollte das Wartungsproblem im lexikalischen Bereich vermieden werden. Ausgehend von einer intellektuellen Analyse von Lexika und Texten (z.B. dem rückläufigen Wörterbuch von MATER) wurde ein Verfahren entwickelt, das zunächst die algorithmische (tabellarische) Identifikation einer Wortklasse (bzw. einer speziellen oder generellen Wortklassenmehrdeutigkeit) erlaubte (so genannte WORTANALYSE). Da die Reduktion einer Wortform auf die Grundform (sowie eine weitere Abtrennung von Suffixen - wie UNG, LICH) von der Ermittlung der aktuellen, d.h. im Text (und bei PASSAT intellektuell zu identifizierenden) belegten Wortklasse abhängig ist, wurde eine syntaktische Analyse angeschlossen, die u.a. ein spezielles Programm zur Disambiguierung syntaktischer Homographen enthält. (Die Ergebnisse der syntaktischen Analyse, v.a. die Strukturbäume, werden allerdings auch zu anderen Zwecken, etwa zum strukturellen Vergleich von Suchanfrage und Dokument bei der Feinrecherche, verwendet. Vgl. Wieland 1978). Wie schon bei der WORTANALYSE sind bei der Reduktion von (Text-)Wortformen auf Grundformen (und Stämme) z.T. Ausnahmen berücksichtigt, v.a. soweit sie die Retrievalverfahren 'stören' können. (D.H. MUSSE, MUSE und MUS müssen/sollten auf verschiedene Stämme / Grundformen abgebildet werden, VERHAELTNISSE und VERHAELTNIS auf die gleiche Zeichenfolge). In der ersten 'Laborversion' von CONDOR sind z.Zt. weite Teile algorithmisch, d.h. über (Spezial)Programme gelöst, dies wird gegenwärtig in einem 'Redesign' modifiziert, etwa wird das System in der Wortform-Grundform-Reduktion (Lemmatisierung) in die Lage versetzt, 'Ausnahmen' auf einfache Weise zu 'lernen', d.h. die 'Regel' zu ändern oder zu erweitern. (Dieses halbautomatische Lernverfahren ist für die WORTANALYSE, d.h. die Identifikation der möglichen syntaktischen Funktionen, bereits implementiert; vgl. Ring 1978.)

Da - im Gegensatz zu PASSAT - in jedem Fall vom System eine (und gegenwärtig: nur eine) Lösung erzielt wird, muss naturgemäß mit Fehlern gerechnet werden, die sich ergeben können aus

- einer falschen Wortanalyse
- einer falschen Disambiguierung von Homographen
- einer falschen/fehlerhaften Grundform- und Stammermittlung

Nach dem derzeitigen Verfahrensstand hängt die Fehlerquote ab von

- dem zugrundegelegten Wortmaterial (bei Klartext - also eingeschränktem Wortschatz - kann die Fehlerquote der Wortanalyse vernachlässigt werden; allerdings ist es z.Zt. über CONDOR nicht möglich - zumindest nicht im linguistischen Teil - ein 'unzulässiges' Wort als solches zu identifizieren (13)).

- der (syntaktischen) Komplexität der Sätze (die linguistische Analyse von CONDOR arbeitet auf Satzebene). Da "längere" Sätze im allgemeinen auch eine höhere Komplexität aufweisen, korreliert die Fehlerquote mit der durchschnittlichen Satzlänge; z.Zt. liegt sie - grob gerechnet - im Bereich der syntaktischen Disambiguierung für CONDOR bei ca. 10%. Diese Zahl erscheint sehr hoch, sie resultiert jedoch aus einer sehr feinen (für weitere Sprachanalysen notwendigen) Differenzierung von Wortklassen (z.B. VERLOREN in "Partizip II als Teil der Verbform", "finites Verb-Präteritum", "Adverb"), der Einfluss dieser "Fehler" im Hinblick auf die Reduktion von Wortformen auf Grundformen und Stämme ist bedeutend niedriger.

Eine semantische Disambiguierung, d.h. eine Differenzierung der Bedeutung von Einzelwörtern, ist auf diesem linguistischen Wege natürlich nicht möglich und auch von CONDOR derzeit nicht beabsichtigt (Ansätze dazu finden sich allerdings in den sprachstatistisch orientierten Teilen von CONDOR).

Der wesentliche Grund für eine 'Vernachlässigung' des Homonymieproblems beruht - abgesehen von der Intention, ein möglichst wartungsfreies System zu entwickeln - auf der Vorstellung, dass beim textuellen Retrieval die Problemstellung des Fragestellers - bezogen auf eine individuelle, gezielte (dialogische) Suche - nur in seltenen Fällen über ein Einzelwort definiert ist, eine 'Vereindeutigung' aber meist über den 'Kontext' der Suchfrage möglich ist. 'Kontext' ist hier nicht (oder nicht nur) in der üblichen Form einer Kombination von Schlüsselwörtern mit Boole'schen Operatoren (UND, ODER, UND NICHT) zu verstehen. Im Mittelpunkt der Konzeption dieses Teils von CONDOR steht die Strategie, die "Suchfrage" als (Mini-) Problembeschreibung in natürlicher Sprache zu verstehen, also eher als "Suchaussage" zu behandeln.

Dabei wird aber nicht allein das gemeinsame Vorkommen mehrerer Wörter wichtig, sonder auch die Relation, in der die Begriffe in einem (Ziel-)Dokument, auch dem 'Suchfragen-Dokument', stehen. Es wird z.Zt. - grob gesagt - davon ausgegangen, dass diejenigen Zieldokumente besonders relevant erscheinen, die die größte "Ähnlichkeit" mit der Suchfrage aufweisen. Lautet z.B. eine Teilstruktur im Suchfragen-'Dokument' (bei CONDOR: Elementarrelation, Wortpaar) BESUCH DES MINISTERS, so werden alle Dokumente, die die gleiche oder ähnliche Struktur aufweisen (z.B. BESUCH/EINES/DIESES MINISTERS, BESUCH VON MINISTER X, BE-

SUCHE DER MINISTER) 'höher' gewichtet als solche, die nicht in dieser textuellen (syntaktischen) Relation auftreten (dazu werden mathematische Funktionen, sog. 'Relevanzfunktionen' herangezogen, die in einer Verknüpfung heuristisch ermittelter Wort-, Struktur- und Stellungsgewichte bestehen; vgl. dazu ausführlich: Wieland 1979).

Einen ähnlichen 'Rang' erhalten Dokumente, in denen statt der syntaktischen Relation ein entsprechendes Kompositum (hier z.B.: MINISTERBESUCH) auftritt. Dies wird ermöglicht durch die automatische Zerlegung zusammengesetzter Wörter mithilfe eines sog. 'Morphembaums' (fürs Deutsche), der inhaltlich einem Morphemlexikon gleichkommt, wobei bei der Systementwicklung - wie schon bei der Entwicklung der Wortklassenidentifikation - entsprechende Lexika, in diesem Falle Morphemlisten des Deutschen, zugrundegelegt sind.

Daneben werden die ermittelten Deskriptoren benutzt, um auf statistischem Weg Dokumenten-Ähnlichkeiten zu berechnen (Clustering), aber auch, um (statistische) Assoziationen zwischen Deskriptoren aufzubauen (vgl. Panyr 1978). Diese Verfahren erlauben es, weitere automatische oder halbautomatische Strategien zum Dokument- Retrieval zu entwickeln.

Gegenwärtig wird die Laborversion von CONDOR ersten Pilottests unterzogen, wobei allerdings die auf linguistischen Kriterien (d.h. der Textdatenverarbeitung) aufbauenden Komponenten nur einen Teil ausmachen. Frühestens Mitte 1979 werden aus der Sicht der Pilotanwender erste grundsätzliche Aussagen über die Wirksamkeit dieser Verfahren vorliegen.

5.3 JUDO

Die Ursprünge des linguistischen Aspekts des CONDOR-Verfahrens gehen auf Kontakte der Projektgruppe mit einer Forschungsstelle an der Universität des Saarlandes zurück, an der in den 60-er Jahren ein Verfahren zur maschinellen Sprachanalyse des Deutschen entwickelt worden war (vgl. Eggers 1969). Im Gegensatz zu CONDOR wurde dort ein streng 'lexikalistischer' Ansatz verfolgt (also etwa die PASSAT- Linie), wenn auch weitere syntaktische Informationen und der Einsatz eines 'Parsers', d.h. eines Analysealgorithmus, weitere (und feinere) Sprachstrukturbeschreibungen ermöglichen sollten. Mit der Einrichtung eines Sonderforschungsbereichs 'Elektronische Sprachforschung' in Saarbrücken zu Anfang der 70-er Jahre wurde zugleich eine Neukonzeption des Verfahrens vorgenommen unter dem Schlagwort der "Automatischen Lemmatisierung", d.h. der automatischen Reduktion von (Text-)Wortformen zu Grundformen einschließlich dem Versuch, auch semantische Mehrdeutigkeiten aufzulösen. Seit 1977 fördert der Bundesminister für Forschung und Technologie in diesem Zusammenhang an der Universität Regensburg das anwendungsorientierte Projekt 'Modellentwicklung eines Software-Systems zur computergestützten Indexierung'. Dieses Projekt hat zum Ziel, den Saarbrücker Parser ('Saarbrücker Automatische Textanalyse' - SATAN -) in einen Indexierungs- und Retrieval-Prozess zu integrieren. Entwickelt und erprobt wird das Gesamtsystem am Beispiel juristischer Dokumente (daher das Projektkürzel JUDO); vorwiegend handelt es sich bei den Testtexten um Normen zum Informationsrecht (z.B. die Paragraphen des Bundesdatenschutzgesetzes) und zum Steuerrecht (eine Teilmenge der bei JURIS gespeicherten Judikate).

Das System JUDO setzt - im Gegensatz zu CONDOR - eine starke lexikalische Komponente voraus. Die Wortformen eines Textes (Satzes) werden über Lexika identifiziert und dadurch in ihren möglichen Funktionen spezifiziert, wobei sie mit syntaktischen und semantischen Informa-

tionen angereichert werden. Voraussetzung für eine erfolgreiche Bearbeitung ist gerade die Ausstattung mit spezifischen Informationen; dennoch sei angemerkt, dass es auch bei diesem Verfahren möglich ist, 'unbekannte' Wörter, d.h. solche, die nicht als Gesamteintrag über ein Lexikon identifiziert wurden, zu verarbeiten: hierzu sind Algorithmen zu Kompositazerlegung und morphologisch-syntaktische Verfahren entwickelt.

Es ist hier nicht möglich, das Verfahren zur Analyse ausführlich darzustellen (14); an dieser Stelle sollen nur einige Parallelen und Unterschiede zu den übrigen Verfahren aufgezeigt werden: Im Unterschied zu CONDOR werden eine größere Zahl syntaktischer Informationen (z.B. bei Substantiven Angaben zu Genus, Numerus, Kasus, bei Verben zu Valenz, Reflexivität) und auch semantische Merkmale und Regeln herangezogen, insbesondere mit dem Ziel der syntaktischen und semantischen Disambiguierung.

Ähnlich zu CONDOR werden syntaktische Strukturen (Nominal- und Verbalphasen, Satz- und Subsatzstrukturen, etwa auf dem Niveau der generativen Transformationsgrammatik) ermittelt, wobei es jedoch zugleich möglich ist, mehrwortige Ausdrücke, so genannte 'Feste Wendungen' (z.B. PERSONENBEZOGENES DATUM, KRAFT AMTES) als 'Einheit' (Deskriptor) zu identifizieren.

Über spezielle Lexika und Thesauri, die allerdings von einem zukünftigen Benutzer zu warten und auszubauen sein werden, können mit Hilfe von JUDO sinnvolle Kompositazerlegungen erfolgen; eine intellektuell gepflegte Relationierung von Begriffen soll beim Retrieval (fakultativ) ein 'natürliches' Befragen ermöglichen. Durch die automatische Sprachanalyse ermittelte syntaktische Strukturen werden unmittelbar zum Aufbau von komplexen Deskriptoren benutzt.

Zur Veranschaulichung der möglichen Verwendbarkeit dieses Verfahrens sei die Deskriptorvergabe etwas ausführlicher behandelt.

Dem Benutzer sollen folgende 'Deskriptorklassen' zum Retrieval zur Verfügung stehen:

i) Normaldeskriptoren

> Normaldeskriptoren sind Textwörter oder Textwortfolgen, die nicht als Teil eines Kompositums oder einer Festen Wendung erschlossen wurden. Sie sind bei Vorliegen von Homonymie (durch Bedeutungsnummern) differenziert, wenn die aktuelle Bedeutung - auf maschinellem Weg oder durch intellektuellen Eingriff - eindeutig bestimmt werden konnte. Alle Normaldeskriptoren können (nach intellektueller lexikalischer Vorarbeit) semantisch / formal relationiert werden. Folgende Verknüpfungen zwischen Normaldeskriptoren sind vorgesehen:

- strenge Synonymie (darunter auch: Rechtschreibvarianten, Vollform-Abkürzung) .
- Oberbegriff - Unterbegriff (und umgekehrt)
- Ganzes-Teil (und umgekehrt)
- Quasi-Synonymie
- Assoziation (semantisches Feld)
- Antonymie

- Derivation (Substantiv - Verb/Adjektiv)
- juristischer Regelungsgegenstand - juristische Regelung im speziellen Rechtsbereich (diese Verknüpfung ist natürlich fachspezifisch)

(ii) Teildeskriptoren

Zur Klasse der Teildeskriptoren gehören alle Deskriptoren, die als Ergebnisse bei der Zerlegung von Komposita oder Festen Wendungen in 'sinnvolle' Elemente ermittelt wurden, z.B. ABGEORDNETER aus BUNDESTAGSABGEORDNETER. Durch diese Differenzierung lassen sich beim Retrieval von Fall zu Fall unterschiedliche Strategien anwenden (Suche mit und ohne Teildeskriptoren).

(iii) Deskriptoren der Klasse 'möglicherweise richtig'

Wenn die maschinelle Textanalyse keine (endgültige) Vereindeutigung eines (homonymen) Deskriptors bewirkt hat und auch keine intellektuelle (interaktive oder posteditive) Disambiguierung erfolgt ist, wird ein Deskriptor in diese Klasse eingeordnet. Auch hier hat der Benutzer die Möglichkeit, über geeignete Strategien diese Deskriptoren in eine Suche einzubeziehen. Über die Deskriptoren dieser Klasse sind auch die semantischen Relationen hergestellt, wie sie für Normaldeskriptoren gelten. (Es handelt sich ja prinzipiell um die gleichen Begriffe, nur dass man nicht weiß, ob die entsprechende Bedeutung in dem betreffenden Dokument auch 'aktualisiert' ist.) Auch wenn eine Bedeutungsvariante eines n-deutigen Wortes ausgeschlossen wurde, werden die verbleibenden n-1 Bedeutungen bei n /= 1 alle dieser Klasse "möglicherweise richtig" zugeordnet, d.h. zu dem jeweiligen Dokument in den noch möglichen Varianten vergeben.

(iv) Komplexe Deskriptoren

(Deskriptoren mit linguistisch-syntaktisch ermittelten Verknüpfungen)

Aus der Vielzahl von möglichen syntaktischen Relationen zwischen Deskriptoren, wie sie aktuell in einem Dokument auftreten, wurden im Rahmen von JUDO zwei herausgesucht, weil sie (aufgrund der vergleichbaren Struktur Fester Wendungen) für besonders interessant gehalten werden:

- die Adjektiv-Substantiv-Relation. Alle Adjektive werden mit dem jeweiligen im Text vorhandenen Substantiv zu einem komplexen Deskriptor zusammengefasst. Beispiel: BESOLDUNGSRECHTLICH adj VORSCHRIFT.

- die Substantiv-Substantiv-Relation. Insbesondere sind hierbei die syntaktischen Verknüpfungen Substantiv + Genitivattribut (VORSCHRIFT gen GESETZ z.B. für 'Vorschrift(en) dieses Gesetzes'), Substantiv mit präpositionalen Anschluss (BEILAGE präp GESETZ z.B. für Beilage zum Gesetz') und Substantiv in konjunktionaler Nebenordnung (GESETZ konj VERORDNUNG z.B. für 'Gesetze und/oder Verordnungen') vorgesehen.

Man kann sich u.a. vorstellen, dass hierbei ein 'Mengenproblem' entsteht, gleichwie man diese Verknüpfungen technisch realisiert. Daher wird sich längerfristig die Frage stellen,

inwieweit man (evtl. über statistische/intellektuelle Verfahren) die Bildung von 'unsinnigen' komplexen Deskriptoren verhindert oder ohne Substanzverlust reduziert.

Es ist denkbar (und z.T. auch vorgesehen), weitere syntaktische Relationen zwischen Deskriptoren (etwa die Zugehörigkeit zum gleichen Teilsatz) zu erproben. Sie alle stellen mehr oder minder Verfeinerungen und Präzisierungen der von STAIRS her bekannten - rein formalen - Relationen (z.B. ADJACENT und SAME) dar. Auf den ersten Blick scheinen jedoch die linguistischen Relationen zu versprechen, die Akzeptabilität des Systems zu erhöhen.

(v) Formaldeskriptoren

Der Vollständigkeit halber sei erwähnt, dass es im Rahmen von JUDO auch möglich sein soll, ein Retrieval ohne Bedeutungsdifferenzierung bei Homonymen durchzuführen. Bei der Verwendung eines mehrdeutigen Deskriptors ohne Bedeutungskennung wird auf alle Dokumente gezeigt, bei denen die Zeichenfolge (Grundform) als potentieller Deskriptor oder als (beliebiges) Zerlegungsergebnis der Dekomposition auftrat. (Damit sollen Benutzerfälle integriert werden, in denen eine Bedeutungsdifferenzierung aufgrund der Fragestellung nicht nötig erscheint). Jedem Formaldeskriptor ist zudem ein Informationsdokument zugeordnet, in dem der Benutzer Aufschluss über die Bedeutungsdifferenzierung (und die daraus resultierenden Normaldeskriptoren) erhält.

Die ersten Erprobungen von JUDO mit konkreten Textdaten (gegenwärtig ist der Aufbau des Gesamtsystems in zwei IR-System-Varianten - TELDOK und GOLEM - unter Verarbeitung des Bundesdatenschutzgesetzes und verwandter Dokumente in Arbeit) haben folgende Erkenntnisse gebracht:

Eine unmittelbare 'Übernahme' von Orginaltexten - ohne eine gewisse, wenn auch kaum ins Gewicht fallende Präkodierung bzw. Anpassung des Parsers an die Textsorte ist kaum möglich. Zumindest der Text des BDSG weist z.T. äußerst komplexe Aufzählungen und Tilgungen auf; 'Sätze' mit über 50 Wörtern sind keine Seltenheit. (Ähnliche - wenn auch anders motivierte - Präkodierungen werden übrigens bei der JURIS-Texterfassung mit PARAT, dem speziellen JURIS-Erfassungssystem, bereits durchgeführt; derartige Präkodierungen sind in der Dokumentationspraxis durchaus üblich und akzeptiert).

Es scheint sich als sinnvoll zu erweisen, die semantische Disambiguierung durch statistische Angaben, d.h. durch Markierungen zur Wahrscheinlichkeit, mit der ein Wort in einer bestimmten Bedeutung in einem (engeren) Fachgebiet auftritt, zu unterstützen. Zumindest für den Bereich des Informationsrechts haben sich entsprechende Verfahren, die bei JUDO erprobt werden, bewährt (15).

Der eigentliche Verfahrenstest steht bei JUDO jedoch - ähnlich zu CONDOR - noch aus, so dass das Jahr 1979 zumindest für die automatische Dokumentation deutscher Texte einen wichtigen Meilenstein darstellt. Nach vielen oft halb leeren Versprechungen (und manchen Zweifeln) sind nunmehr jedenfalls technologisch die Voraussetzungen für eine Erprobung linguistischer Verfahren geschaffen. Nach der Ernüchterung zum Ende der 60-er Jahre tut man auch weiterhin gut daran, in diesem Bereich nicht an perfekte Lösungen zu glauben. Ei-

nige Fragen, etwa die nach dem (Rechen-)Zeitaufwand oder dem Arbeitsaufwand zur Wartung der Systeme und Pflege etwaiger Lexika, denen sich der Anwender selbst bei der prinzipiellen Eignung der Verfahren gegenüber sieht, bleiben zunächst noch offen. Hierüber wird man (vielleicht unter den dann erfahrungsgemäß gewandelten technologischen Voraussetzungen: man denke nur an die Entwicklungsmöglichkeiten im Bereich der Mikroprozessor- und Parallelprozessor-Technik; vgl. z.B. Bell 1976) nachdenken müssen, wenn die 'Ansätze' zu einer realistischen computergestützten Indexierung auf linguistischer Grundlage sich als brauchbar erwiesen haben.

ANMERKUNGEN:

1) Ein gewisses terminologisches Problem bei diesem Begriff resultiert daraus, dass unter Freitextverfahren gelegentlich auch die Methode der Extraktion von Textwörtern (vgl. z.B. STAIRS) verstanden wird.

2) Vgl. zum Begriff der Klartextverarbeitung und zur Problematik auch Wingert 1978. Das Problem Sprachsynthese sei in diesem Zusammenhang ausgeklammert.

3) Ungeklärt - aber vielfach befürchtet - ist das Problem der Verarmung der natürlichen Sprache durch Gewöhnung an solche Restriktionen. Auf diese Fragestellung kann hier jedoch nicht weiter eingegangen werden.

4) Inzwischen ist eine der natürlichsprachigen Darstellung wohl näherkommende Variante (TITUS-4) in Arbeit; obwohl darüber aus kommerziellen Gründen kaum etwas bekannt ist, darf man annehmen, dass noch erhebliche syntaktisch-semantische Restriktionen vorliegen werden, dagegen die morphologische Ausdrucksbreite stark verbessert sein wird.

5) Ein kleines Beispiel für ein derartiges Umdenken aus der Vergangenheit sei erwähnt: noch zu Anfang/Mitte der 70-er Jahre haben nur wenige Anwender (auch im Textverarbeitungsbereich) die Differenzierung nach Groß-Klein-Schreibung bei Computerausdrucken (übrigens auch in der Internverarbeitung, vgl. die üblichen Kartenlocher) Bedeutung zugemessen. Dies war für Verfahren wie CONDOR ein wesentlicher Grund, auf derartige Unterscheidungskriterien, die für die Sprachanalyse des Deutschen doch von gewisser Bedeutung sind, zu verzichten. Heute ist ein Rechenzentrum ohne Drucker mit Groß-Klein-Schreibung, bald wohl auch ein entsprechendes Terminal, kaum noch denkbar. In der maschinellen Textverarbeitung ist das gewohnte Druckbild alltäglich.

6) Dies sollte nicht ausschließen, dass bekannte und bewährte Verfahren der Suchanfrage weiterhin bestehen bleiben; daneben werden sich also auch andere benutzerfreundliche Verfahren (Menütechnik, graphischer Dialog) behaupten und fortentwickeln.

7) Dies soll einmal zur Anschaulichkeit beitragen; die in diesem Rahmen nicht mögliche Ausweitung hätte auch intensiverer Beschreibungen weiterer Systeme bedurft und insbesondere Verfahren der maschinellen Sprachübersetzung einbeziehen müssen. Hierzu sei in diesem Zusammenhang auf die ausführlichen Literaturangaben in Bruderer 1978, Sparck Jones/Kay 1973, Hutchins 1975 und Kuhlen 1977 verwiesen.

Für allgemeine Probleme des Automatic Indexing sei auf Sparck-Jones 1974 hingewiesen. Unter dem Aspekt der Verwendung linguistischer Verfahren ist das Thema bei Hutchins 1970 behandelt. Zu den Arbeiten im deutschsprachigen Raum, die im Folgenden nicht näher behandelt werden können, rechnen u.a. Müller 1973, Lustig 1974, Braun/Schwind 1975, Jaene/Seelbach 1975, Braun 1976 und Schott 1978. Vgl. auch den Beitrag von Lustig in diesem Band.

8) Hier soll allein der "linguistische Ansatz" von CONDOR von Interesse sein (vgl. dazu auch Wieland 1978), Wieland/Haller 1978. Zum Gesamtkonzept von CONDOR, das unter Einbeziehung von GOLEM-Funktionen als allgemeines Information-Retrieval-System verstanden werden soll, vgl. Banerjee, im 2. Band dieses Sammelwerks.

9) JUDO ist ein Akronym für 'Juristische Dokumentanalyse'.

10) Das Saarbrücker System ist grundsätzlich für die Anwendung auf mehrere Sprachen ausgerichtet; in JUDO wird v.a. die deutschsprachige Variante von SATAN (Saarbrücker automatische Text-Analyse) verwendet, vgl. dazu insbesondere Zimmermann 1978a und b.

11) Vgl. den Beitrag von Löhlein in diesem Band

12) Zum Problem dieses "Rauschens" vgl. Peuzner 1973

13) Dies ist allerdings im weiteren Indexierungsprozess zu realisieren.

14) Vgl. dazu das Handbuch der Automatischen Lemmatisierung (1978) und die Kurzbeschreibung in Maas 1979.

15) Vgl. die Untersuchung von Kopelent 1978.

LITERATUR:

Banerjee, N. 1979: CONDOR - Modell eines DB/IR-Systems. In Band 2 dieser Reihe.

Bell, C.L.M. 1976: Minicomputer Retrieval System with Automatic Root Finding and Roling Facilities. In: Program News of Computers in British University Libraries 10 (1976) 14-27

Berry-Rogghe, G.L.; Wulz, H. 1976: The Design of PLIDIS, a Problem Solving Information System with German as Query Language Institut für Deutsche Sprache, Dezember 1976

Billmeier, R.: Ein pragmatisch orientiertes linguistisches Analyseverfahren zur Automatischen Strukturerkennung für deutsche Sätze: Anwendung eines ATN-Parsers in einem Informationssystem.

Braun, S. 1976: Linguistically Based Methods for Indexing and Thesaurus Construction for Information Systems. (Vortrag für die IFIP Infopol Konferenz 1976 Warschau)

Braun, S.; Schwind, C. 1975: Automatic, Semantics-based Indexing of Natural Language Texts for Information Retrieval Systems, TU München, Report Nr. 7505, 153 (1975)

Bruderer, H.E. 1978: Handbuch der maschinellen und maschinen-unterstützten Sprachübersetzung. München 1978

Eggers, H. et al. 1969: Elektronische Syntaxanalyse der deutschen Gegenwartssprache, Tübingen 1969.

Hutchins, W.I. 1970: Linguistic Processes in the Indexing and Retrieval of Documents. In: Computer Physics Communications 1970, 29-64.

Hutchins, W.I. 1975: Languages of Indexing and Classification. London 1975

Jaene, H.; Seelbach, D. 1975: Maschinelle Extraktion von zusammengesetzten Ausdrücken aus englischen Fachtexten. Berlin 1975

Kopelent, J. 1978: Zur Auflösung semantischer Mehrdeutigkeiten bei Verben in juristischen Fachtexten. Regensburg 1978 (Magisterarbeit)

Kuhlen, R. 1977: Experimentelle Morphologie in der Informationswissenschaft. München 1977

Lustig, G. 1974: Probleme der Wörterbuchentwicklung für das automatische Indexing und Retrieval. In: Nachrichten für Dokumentation 25 (1974) S. 50-54

Maas, H.D.: Das Saarbrücker Verfahren zur Automatischen Textanalyse. In: Sprache und Datenverarbeitung 3 (1978) (im Erscheinen)

Müller, S. 1973: Untersuchung zum Inhaltserschließungssystem TELDEX. Konstanz 1973

Panyr, J. 1978: STEINADLER - Ein Verfahren zur automatischen Deskribierung und zur automatischen thematischen Dokumentenklassifikation. In: Nachrichten für Dokumentation 1978,

Perzner, B.R. 1973: The Effect of Indiscriminated Homonymy on Noise Level. In: Automatic Documentation and Mathematical Linguistics 7 (1973) 19-27

Ring, H. 1978: PELIKAN - ein Lernsystem für linguistische Klassifikationsalgorithmen. In: Nachr. f. Dok. 29 (1978) 224-226

Schott, G. 1978: Automatische Kompositazerlegung mit einem Minimalwörterbuch zur Informationsgewinnung aus beliebigen Fachtexten. In: Klartextverarbeitung (ed. F. Wingert) Berlin 1978, 32-43

Seelbach, D. 1975: Computerlinguistik und Dokumentation. Key Phrases in Dokumentationsprozessen. München 1975

Sparck Jones, K. 1974: Automatic Indexing. In: Journal of Documentation 4 (1974) 393-432

Sparck Jones, K.; Kay, M. 1973: Linguistics and Information Science (Dt. Ausgabe u.d.T. Linguistik und Informationswissenschaft) N.Y./London 1973

Wieland, U. 1978: Linguistische Analyse im Projekt CONDOR. In: Klartextverarbeitung (ed. F. Wingert), Berlin 1978, 21-29

Wieland, U.; Haller, J. 1978: Die Erschließung natürlichsprachlicher Information im Informationssystem CONDOR. Nachrichten für Dokumentation 29 (1978), 177

Wieland, U. 1979: Recherche auf der Basis einer syntaxorientierten maschinellen Sprachanalyse (Diss. Regensburg, im Erscheinen)

Wingert, F. 1978: Klartextverarbeitung in der Medizin. In: Klartextverarbeitung (ed. F. Wingert), Berlin 1978, 1-20

Zimmermann, H. 1978 a: Probleme der automatischen Indexierung von Fachtexten am Beispiel juristischer Dokumente. In: Klartextverarbeitung (ed. F. Wingert), Berlin 1978, 112-121

Zimmermann, H. 1978 b: Automatische Textanalyse und Indexierung. In: Kolloquium zur Lage der linguistischen Datenverarbeitung (ed. D. Krallmann) Essen 1978, 20-33

Zingel H.J.: Experiences with TITUS 2. IN: DSP: Deutscher Dokumentartag 1977 (3.10.-7.10.1977) In Saarbrücken. Minden: Verlag Dokumentation 1978, S. 311-330.