In: F. Wingert (ed., 1978): Klartextverarbeitung. Berlin, 112-121

PROBLEME DER AUTOMATISCHEN INDEXIERUNG VON FACHTEXTEN AM BEISPIEL JURISTISCHER DOKUMENTE

H. Zimmermann

"Ein Computer versteht keine Ironie. Er verfügt nicht über das entsprechende Weltwissen, Situationswissen, Textwissen. Man müsste es ihm vermitteln. Ob das gelingt, ist fraglich..." Dieses Zitat aus einer noch druckfrischen Veröffentlichung [3] spiegelt sehr gut die unbefriedigende Situation wider. Und das nach nunmehr schon über 25-jährigem Bemühen um die Verwendung linguistischer Methoden in der maschinellen Textverarbeitung, grob wiederzugeben mit den Schlagwörtern: Automatische Sprachübersetzung, Maschinelles Indexing und Automatisches Abstracting.

Die Entwicklung der theoretischen Linguistik hat sicherlich in den letzten beiden Jahrzehnten nicht stagniert. Formale Sprachbeschreibungsmodelle, wie etwa die Generative Transformationsgrammatik, wurden auch als besonders "computeroperabel" begrüßt (der Begründer dieses Grammatikmodells, CHOMSKY, hat selbst großen Anteil an der Entwicklung einer Typisierung der formalen Sprachen in der Informatik). Auch die z.T. unterschiedlichen Weiterentwicklungen, wie die "Generative Semantik" oder die (Tiefen-)"Kasustheorie", erscheinen wegen ihrer Formalisierbarkeit EDV-praktikabel. Im Zusammenhang mit sprachlich orientierten Modellen und Systemen der "Artificial Intelligence-Forschung" sind ebenfalls, vor allem in den letzten zehn Jahren, genügend Systeme bekannt geworden, die "Sprachverstehen" simulieren bzw. demonstrieren wollen. Im Dokumentations- und Informationsbereich praktisch einsetzbare Systeme zur "Textverarbeitung" auf einem genügend hohen Sprachverstehensniveau liegen aber bislang nicht vor. Ohne Zweifel - dies unterstreicht auch die zu Anfang erwähnte Veröffentlichung - ist dafür die Komplexität der natürlichen Sprache (d.h. der normalen sprachlichen Äußerungen) verantwortlich. Auch eine Beschränkung auf Fachtexte (z.B. mit Einschränkungen im Wortschatz und im Satzbau) hilft da kaum weiter, vor allem dann nicht, wenn die zu analysierenden Äußerungen - wie in der Regel üblich - im Hinblick auf die Kommunikation Mensch-Mensch entstanden sind.

Die technische Entwicklung gerade der letzten Jahre (z.B. Verfügbarkeit großer Speichermedien, maschineller Satz, optische Leser,...) ermöglicht es inzwischen, große Textmengen mit - im Vergleich zu früher - geringem Aufwand maschinell zu verarbeiten. Während die Erfassungskosten z.B. noch vor zehn Jahren ein wichtiges ökonomisches Argument gegen eine maschinelle Übersetzung bildeten (vgl. den ALPAC-Report), entstehen computeroperable Textdaten heute zum Teil schon als Abfallware (etwa bei Verlagsprodukten). Dies gilt auch für maschinelle Informationssysteme, in denen textuelle Informationen (d.h. Klartext) neben den formalen (retrievalfähigen) Daten als Hintergrundinformation gespeichert sind.

Während die maschinelle Textverarbeitung in Bereichen wie dem einer qualitativ guten maschinellen Übersetzung, eines dem humanen vergleichbaren automatischen Abstracting oder eines ein grösseres Wissensgebiet (Midi- oder Makrowelt) umfassenden Fact-Retrieval aus den genannten Komplexitätsgründen derzeit nicht praktikabel erscheint - allenfalls ist hierbei Computer-Unterstützung denkbar - , ist gegenwärtig das automatische Indexing am ehesten für den praktischen Einsatz geeignet. Nach den bisherigen Erfahrungen, die auch bei Vergleichen innerhalb und mit der intellektuellen Indexierung gesammelt wurden, sind einer maschinellen Indexierung, d.h. ei-

ner Vergabe von Deskriptoren zu einem Dokument aufgrund einer maschinellen Analyse des in einem Text verwendeten Wortmaterials, am ehesten gewisse Chancen einzuräumen.

In der folgenden Betrachtung soll nicht auf statistisch orientierte Indexierungsmethoden eingegangen werden. Die Ausführungen beschränken sich vielmehr im Wesentlichen auf linguistische Fragen, die entweder beim Ermitteln potentieller Deskriptoren aus dem Textmaterial entstehen oder sich beim Retrieval mit den maschinell ermittelten Wörtern ergeben. Die Beschreibung orientiert sich dabei in den meisten Fällen an Beispielen aus dem Bereich der Rechtsdokumentation, insbesondere zum Sozial- und Steuerrecht. Dazu werden Erfahrungen eingebracht, die an in Deutschland eingesetzten Systemen zur automatischen Indexierung gewonnen wurden. Das Juristische Informationssystem der Bundesregierung (JURIS) verwendet beispielsweise (bislang) das System GOLEM/PASSAT von Siemens, wobei PASSAT die "linguistische" Analyse bei der Indexierung zukommt. Die Steuerrechtsdokumentation der DATEV arbeitet mit dem System STAIRS von IBM. Damit sind zugleich zwei prototypische Systeme genannt, die eine brauchbar-praktische Ausnutzung textueller Informationen für das Dokument-Retrieval ermöglichen sollen. Da diese Systeme kommerziell vertrieben werden, allgemein zugänglich und auch relativ gut dokumentiert sind, sollen hier nur einige für unser Problem relevante Systemkomponenten und Funktionen beschrieben werden, wobei (vor allem zu STAIRS) Teile der Retrievalkomponente einbezogen sind.

Im System STAIRS (IBM) bilden die im aktuellen Text vorkommenden Wortformen die wesentliche Retrievalgrundlage. Irgendwelche Reduktionen (auf Grundformen z.B.) finden bei der Indexierung nicht statt. Notiert wird allerdings die jeweilige Belegstelle einer Wortform (wie Abschnitt, Satz, Wortnummer im Satz). In dem so entstandenen Wörterbuch lassen sich die Wortformen zusätzlich intellektuell synonym verknüpfen.

Zum Retrieval stehen damit alle Wortformen der Texte, Belegstellen- und Häufigkeitsangaben sowie die Synonym-Markierung zur Verfügung. Da Wortformen in flektierenden Sprachen unterschiedliche Realisierungen eines Begriffs darstellen, wurde zum Retrieval eine Maskierungsfunktion zur Verfügung gestellt (Wortendemaskierung mit fakultativer Angabe der Maximallänge des Restwortes), so dass Flexionsformen zusammengefasst werden können:

(1) ERGEBNIS$3 -- ERGEBNIS; ERGEBNISSE; ERGEBNISSES; ERGEBNISSEN
ERGEBNIS$ -- ..., ERGEBNISKONTROLLE, ...

Dies setzte (bislang) voraus, dass ein Bearbeiter sich die möglichen morphologischen Realisationen eines Begriffes vergegenwärtigen musste, um sicherzugehen, dass alle Belege nachgewiesen wurden. Zusätzlich mussten Unregelmäßigkeiten im Wortstamm berücksichtigt werden

(2) GEHEN -- GEH$, GING$, GEGANGEN$ (Ablaut)
HAUS -- HA$ bzw. HAUS$, HAEUSER$ ...

Dieses Problem soll durch einen inzwischen entwickelten Retrieval-Baustein (STAIRS-TLS) behoben werden, bei dem in der Suchfrage nur noch die Grundformen angegeben werden müssen, aus denen dann automatisch die potentiellen Wortformen generiert (bzw. nach der Recherche - für den Benutzer unsichtbar - selektiert) werden.

(3) HAUS -- HAUS, HAUSES, HAUSE, HAEUSER, HAEUSERN

Weitere Probleme ergeben sich aus den möglichen unterschiedlichen Schreibweisen einer Wortform (sog. Wortformenalternanten):

(4)	PHOTO	-- FOTO
	PHOTOGRAPH	-- FOTOGRAPH
	BFH	-- BUNDESFINANZHOF (Abkürzung)
	GrEStG	-- GRUNDERWERBSTEUERGESETZ
	NIESSBRAUCHRECHT	-- NIESSBRAUCHSRECHT (Fugen-S)
	EINNAHMEN-UEBERSCHUSS	-- EINNAHMENUEBERSCHUSS (Wortbindestrich)

Diese Schreibvarianten können bei STAIRS allein mit Hilfe der Synonymie-Relation abgedeckt werden. Dadurch wird diese Funktion allerdings im Hinblick auf ihre eigentlich semantische Aufgabe - nämlich bedeutungsmässig eng verwandte Begriffe miteinander zu verknüpfen - :

(5)	EHEGATTEN	-- EHELEUTE (Synonymie)
	AUFZUG	-- FAHRSTUHL

zumindest partiell umfunktioniert.

Im System GOLEM/PASSAT (Siemens) übernimmt - wie schon erwähnt - PASSAT die Indexierung. Grundlage ist eine 'Vergleichswortliste', d.h. ein Wörterbuch, das alle in den zu bearbeitenden Texten (Dokumenten) vorkommenden Wortstämme (auch - bei unregelmäßigen Flexionsformen - Pseudostämme) enthalten muss ('unbekannte Stämme' sind nicht zulässig); diesen Graphemfolgen sind Angaben über zulässige Flexionsendungen, Fugenzeichen bei der Komposition und eine 'Assoziationsmatrix' mitzugeben (wobei die Matrix im Allgemeinen - z.B. bei JURIS - allerdings nur zur Identifikation von Stoppwörtern, d.h. als Deskriptoren zu eliminierenden Wörtern, verwendet wird).

Außerhalb von PASSAT (im GOLEM-Thesaurus) sind dann noch verschiedene Verknüpfungen von Deskriptoren (z.B. Synonymie) möglich. In PASSAT erfolgt also schon bei der Indexierung eine Reduktion von Wortformen auf Grundformen, die Verträglichkeitsmarkierung für die Komposition wird von einem Systembaustein zur Zerlegung von Komposita verwendet:

(6)	RECHTSCHUTZ	-- RECHT + SCHUTZ (?)
	KNAPPSCHAFTSRENTE	-- KNAPPSCHAFT + RENTE

Erwähnt werden muss allerdings, dass die Kompositazerlegung - da keine semantischen Restriktionen eingebracht werden können - zu Ballast in der Deskribierung eines Dokuments führen kann:

(7)	RINDERSTALL	-- RIND + ERST + STALL + ALL
	ABWEICHEN	-- (AB) + WEICHEN (!)
	QUARTALSENDE	-- QUARTAL + SENDEN + ENDE
	PROZESSHANDLUNGEN	-- PROZESS + HANDLUNG + HAND + LUNGE

Während - wie die gegenwärtigen Systeme ausweisen - das Problem der Wortformen-Wortstamm-Reduktionen halbwegs gelöst erscheint, sind maschinelle Verfahren mit weitergehenden linguistischen Analysen bislang nicht produktionsreif und gegenwärtig allenfalls in der Vorphase für Pilotanwendungen einsetzbar. Im Folgenden soll anhand zweier ebenfalls in der Vorgehensweise unterschiedlicher Systeme, die in Deutschland in Entwicklung sind, auf die für die nähere Zukunft (ca. 1980) zu erwartenden Möglichkeiten hingewiesen werden, einige der bislang noch unberücksichtigten Komponenten der automatischen Sprachanalyse in ein produktionsreifes Indexierungs- und Retrievalverfahren zu integrieren. Dies betrifft vor allem die Einbeziehung des Kontexts, insbesondere im Rahmen der Syntax. Das System CONDOR [2] hat zum Ziel, im Bereich der automatischen Indexierung vor allem folgende Probleme zu bewältigen (die zum Teil schon angesprochen wurden):

a) die Flexions(endungs)bereinigung ('Lemmatisierung'),
b) die Reduktion von syntaktisch mehrfunktionalen Wortformen aufgrund des Kontexts im Satz (Homographenreduktion).

CONDOR wird daneben ein vollständiges Retrieval-System umfassen einschließlich Recherchefunktionen für formatgebundenes Retrieval. Hier beschränke ich mich auf den Teil der linguistischen Analyse, der größere Möglichkeiten für einen praktischen Einsatz bieten wird [4]

(8) Homographie:

EHE	--	Konjunktion	(... ehe er kommt ...)
		Substantiv	(... in der Ehe ...)
EINIGEN	--	Indefinitpronomen	(... mit einigen Gläsern)
		Adjektiv	(... mit einigen Freunden)
		Verb	(... wir einigen uns)
LAUTEN	--	Verb (LAUTEN)	Die Fragen lauten ...
		Subst. (LAUT)	Mit schrillen Lauten ist ...
		Adj. (LAUT)	Mit lauten Schreien ...
BILLIGEN	--	Verb (BILLIGEN)	Sie billigen es ...
		Adj. (BILLIG)	Die billigen Apfel ...

c) Kompositazerlegung (sinnvolle Einheiten).

Ausgangspunkt ist eine so genannte 'Wortanalyse', in der die potentiellen syntaktischen Funktionen einer Wortform ermittelt werden sollen. Dabei wird ein Funktionswörterbuch herangezogen, das alle in einer Sprache vorkommenden Partikelwörter (wie Konjunktionen, Präpositionen, Pronomina) als geschlossene Liste enthält. Um eine größtmögliche Wartungsfreiheit zu gewährleisten, werden die Hauptwortklassen Verb, Substantiv, Adverb und Adjektiv algorithmisch (über Graphemkombinationen) ermittelt. Die sich an die Wortanalyse anschließende 'Homographenanalyse' versucht in einer für ein relativ zuverlässiges Retrieval erforderlichen Exaktheit eine Vereindeutigung der syntaktischen Funktion eines potentiell mehrdeutigen Wortes aufgrund der Wortumgebung im Satz zu erreichen. Darauf baut dann einerseits eine syntaktische Strukturbeschreibung, andererseits die 'Lemmatisierung' auf. Diese Lemmatisierung (die durch eine zusätz-

liche Suffixanalyse auch wortklassenunabhängige Lemmata zu bilden erlaubt) ist eine wesentliche Voraussetzung für die Vergabe der Deskriptoren.

Es liegt auf der Hand, dass mit einer syntaktischen Analyse bei weitem nicht alle Probleme einer maschinellen Textaufbereitung gelöst werden können. Dennoch stellt sie meines Erachtens einen ersten Schritt zur Behandlung der für ein intellektuell gesteuertes Retrieval so wichtigen semantischen Verknüpfung von Wörtern dar. Bei CONDOR wird hierzu - auf der Basis der lemmatisierten Einheiten (Wörter) - maschinell auf statistisch-distributioneller Grundlage eine Art 'semantisches Netz' erstellt, das dazu dient, dem Benutzer Hilfen bei der Auswahl von Deskriptoren während des Retrieval-Prozesses anzubieten [2, 4] .

Ein wesentliches Merkmal von CONDOR bildet der weitgehende Verzicht auf ein informationsreiches Lexikon der deskriptorfähigen Wörter und damit auch die Vermeidung einer aufwendigen zusätzlichen intellektuellen Vergabe lexikalischer Informationen. Dies soll zum Teil durch 'Lernalgorithmen', die auf der textuellen syntaktischen und semantischen Information (Netz) aufbauen, automatisch (wartungsfreundlich) ausgeglichen werden. Beispielsweise wird auf explizite Angaben zum Valenzrahmen von Verben und zur semantischen Subklassifizierung von Nomina oder Adjektiven verzichtet:

(9)

BEZIEHEN --	Nom (+hum) Akk (-abs, -bel)
	OMA BEZIEHT EINE NEUE WOHNUNG
BEZIEHEN2 --	Nom (+hum) Akk (+abs)
	ER BEZIEHT EINEN NEUEN STANDPUNKT
BEZIEHEN3 --	Nom (+hum) Akk (-abs, -bel) 'MIT' (-abs, -bel)
	SIE BEZIEHT DAS BETT MIT LEINENTUECHERN
BEZIEHEN4 --	Nom (+hum) Akk ('sich') 'auf' (±abs, ±hum)
	ICH BEZIEHE MICH AUF DICH / AUF DAS VERSPRECHEN
VERMESSEN --	Nom (+hum) Akk (-abs, -bei)
	DER SCHREINER VERMISST DEN TISCH
VERMISSEN --	Nom (+bel) Akk (±abs, ±bei ...)
	DER HUND VERMISST SEINEN HERRN
	DAS KIND VERMISST ZAERTLICHKEIT
KIND	-- +hum ...
ZAERTLICHKEIT	-- +abs ...
VERSPRECHEN	-- +abs ...

Damit werden - zumindest bislang - Merkmale vernachlässigt, die gerade im Hinblick auf eine syntaktische und semantische Disambiguierung nützlich sein können. Allerdings geht man bei CONDOR davon aus, solche Deskriptoren ggf. auf automatischem Weg - mit Hilfe des 'semantischen Netzes' - ermitteln zu können.

Das zweite Verfahren, das hier vorgestellt werden soll - die 'Saarbrücker automatische Textanalyse' (SATAN) - versucht die syntaktisch-semantischen Vereindeutigungen mit Hilfe eines entsprechend vorstrukturierten maschinellen Lexikons zu erreichen. Diesem Konzept liegt also die Annahme zugrunde, dass sich das allgemein- und auch das fachsprachliche Wissen, vor allem was die Verknüpfbarkeit von Wörtern in einem Satz/Text anbelangt, zum Teil in einem Lexikon durch Merkmale repräsentieren (und damit intellektuell präkodieren) lässt. Diese Informationen

sollen nicht zur semantischen Interpretation eines Satzes/Textes verwendet werden - dazu sind sie wohl noch zu sehr an der Oberfläche orientiert - , sondern werden benutzt zur Verbesserung der syntaktisch-semantischen Disambiguierung von Textwörtern.

Bei SATAN wird die Flexionskomponente - ähnlich PASSAT - mit Hilfe eines Stammwörterbuches und der Angabe der zulässigen Endungen behandelt. Allerdings dienen diese Merkmale auch dazu, Angaben zur möglichen Konjugations- oder Deklinationsform eines Wortes zu erhalten:

(10)
HAUS	-- Nom/Dat/Akk Singular
HAEUSER	-- Nom/Gen/Akk Plural
DAS HAUS	-- Nom/Akk Singular
Der HAEUSER	-- Gen Plural

Diese Informationen werden beim Parsing zur Kontrolle von Kongruenzen herangezogen. Die syntaktische Analyse berücksichtigt in erster Linie Wortstellungs- und Valenzrestriktionen, die sich zunächst auf die syntaktische Oberfläche beziehen. In einem zweiten Schritt werden dann feinere Merkmale herangezogen (vgl. (9)), in Zukunft sollen auch Relationen zwischen Wörtern (z.B. Hypernymie-Relation) erstellt und in den Beschreibungs- und Disambiguierungsprozess eingebracht werden.

In Regensburg soll nun versucht werden, das Saarbrücker Verfahren in einem automatischen Indexierungsverfahren im Bereich der Rechtsdokumentation zu erproben. Ausgangsmaterial ist das Steuerrecht (Judikate) sowie eine Sammlung von Texten zum EDV-Recht (Normen bzw. Normenentwürfe). Bei den Voruntersuchungen hat sich bereits gezeigt, wie notwendig in diesem Zusammenhang die Ermittlung mehrwortiger Ausdrücke ist.

(11)
KRAFT	-- IN/AUSSER KRAFT TRETEN
BILLIGEN	-- NACH BILLIGEM ERMESSEN
BEHALTEN	-- IM AUGE BEHALTEN
WILLEN	-- WILLENS SEIN / UM ... WILLEN / ZU WILLEN SEIN ...

Obgleich das Saarbrücker System dazu ein Beschreibungs- und maschinelles Erkennungsverfahren aufweist, hat sich herausgestellt, dass die dort erfassten rund 5000 allgemeinsprachlichen 'festen Wendungen' für die Fachtextindexierung nicht ausreichen. Es gilt daher, diese Listen um die fachtextbezogenen Daten zu erweitern.

Des Weiteren soll versucht werden, bei der Indexierung nicht - wie noch im Saarbrücker Verfahren vorgesehen - die Wortklassengrenze bei der Wortformenzuordnung beizubehalten, sondern - soweit sinnvoll - z.B. eine Abbildung von Verben auf entsprechende Nomina u.ä. durchzuführen. Erfahrungen bei der Untersuchung typischer Retrieval-Fragen bei JURIS haben nämlich deutlich werden lassen, dass die Nomina bei Suchfragen bevorzugt werden. Dagegen zeigt die Untersuchung der beiden juristischen Fachtext-Bereiche, dass die Verbformen (und die daraus abgeleiteten Partizipien) einen durchaus relevanten Anteil im Text besitzen. Ein großes Problem stellt die automatische Zerlegung der Komposita in für ein Retrieval relevante Bestandteile dar. Dazu gibt es bislang keine zufriedenstellenden, praktisch erprobten automatischen Verfahren (obwohl Versuche dazu an den verschiedensten Stellen vorliegen). Der lexikalische Ansatz erlaubt es nun, solche sprachlich bzw. fachsprachlich sinnvollen Zerlegungen bereits im Lexikon vorzugeben

(Ein Teilproblem - die Bewältigung der 'Augenblickskomposita' bleibt dadurch natürlich noch offen. Hier können provisorische Algorithmen eingreifen bzw. die Lexikonerweiterung unterstützen). Normalerweise setzt die Markierung von Kompositionselementen einen intellektuellen Aufwand voraus, der sich aber durch bessere Retrieval-Ergebnisse rechtfertigt.

Man wird sich in der Zukunft wohl verstärkt der Frage widmen müssen, inwieweit - und in welchen Bereichen - eine Automatisierung in der Klartextdokumentation noch intellektuellen Aufwand sowohl des Sachbearbeiters/Fachmanns als auch (etwa bei der Präkodierung der Texte) eines 'Textbearbeiters' oder 'Textorganisators' erfordert, der um die kritischen (auch linguistisch problematischen) Systemkomponenten weiß und entsprechende systemunterstützende Markierungen vornimmt.

Es bleibt dabei abzuwarten, ob sich der sicherlich wartungsfreundliche prozedurale Ansatz von CONDOR in der Praxis bewähren wird und wo gegebenenfalls seine Grenzen sein werden. In bezug auf eine ökonomische Betrachtung wird der lexikalistische Ansatz - wie im Saarbrücker Verfahren - sicher noch manche Nachteile aufweisen, doch bildet dieser Weg meines Erachtens gegenwärtig eine solide Basis für ein zukunftsorientiertes Indexierungssystem. Es ist vielleicht zu erwarten, dass sich auf längere Sicht ein Kompromiss zwischen (automatisch lernenden) prozedurorientierten und lexikonorientierten Verfahren ergeben wird: soviel intellektueller Aufwand wie nötig, soviel Automatisierung wie möglich. Bei der Komplexität der natürlichen Sprache werden wir uns wohl trotz einiger Möglichkeiten in der automatischen Sprachverarbeitung noch für einige Zeit darauf einrichten müssen, den Computer durch extensive Präkodierung der Texte, lexikalische Information und gegebenenfalls Interaktionen während der Verarbeitung zu unterstützen, wenn wir auch nur halbwegs gute Ergebnisse erwarten wollen.
Daher sei zum Abschluss noch einmal auf das Eingangszitat verwiesen und auch seine Fortsetzung zitiert: "Ein Computer versteht keine Ironie. Er verfügt auch nicht über das entsprechende Weltwissen, Situationswissen, Textwissen. Man müsste es ihm vermitteln. Ob das gelingt, ist fraglich, aber nicht unwahrscheinlicher, als noch vor 5o Jahren die heutigen Computer erschienen wären".

LITERATUR

[1] AMMON, R. v.: Erste Ergebnisse einer Analyse zur Deskribierung juristischer Dokumente mit Hilfe von PASSAT. Arbeitspapier JUDO-A-03. (Regensburg Nov. 1976).

[2] BANERJEE, N.: Verwendung der natürlichen Sprache im Dialogverkehr mit Informationssystemen. In: Matthöfer, H. (Hrsg.): Forschung aktuell. Datenverarbeitung (1976).

[3] KLEIN, W.: Organisation des Wissens durch die Sprache. IBM-Nachrichten 27 (1977) 11-17.

[4] WIELAND, W.: Syntaktische und semantische Analyse natürlichsprachlicher Sätze im Projekt CONDOR. (in diesem Band).