In: I. Dahlberg und M. Schader (Hrsg. 1983): Automatisierung in der Klassifikation. Frankfurt/Main: Indeks Verlag. Studien zur Klassifikaton, Bd.13 (1983) S.14-32

Harald H Zimmermann

AUTOMATISCHE INDEXIERUNG - ENTWICKLUNG UND PERSPEKTIVEN

Gliederung

1. Eingrenzung und Einordnung des Themas

Angesichts der bestehenden Vielfalt von Methoden und Modellen zur Anwendung des Computers für das Erschließen und Wiederfinden von Texten erscheint es wenig sinnvoll, eine gleichsam endgültige, präzise Definition für den Themenbereich der Automatischen Indexierung und verwandter Gebiete (wie z.B. der intellektuellen oder computergestützten Indexierung) zu geben. Dies gilt besonders, wenn man die anstehende Entwicklung der Mikroprozessortechnik vor Augen hat. Formulierungen wie "Inhaltsanalyse . . . ohne maschinelle Hilfe" (bei der Definition von intellektueller Indexierung, DIN 31623) werden zumindest relativiert, wenn die Verwendung von Bürosystemen mit Textverarbeitung und -archivierung tägliche Praxis zu werden beginnt.

Wie bereits die Grenzen zwischen nicht-computergestützter, computerunterstützter und automatischer Inhaltsanalyse verwischen, so gilt Ähnliches bezüglich der Inhalte, d.h. auch der Art und Tiefe der Analyse und Erschließung des zu bearbeitenden Gegenstandes selbst. Für den Bereich der *Text*erschließung, d.h. die Bearbeitung *natürlichsprachiger* Dokumente, trifft dies in besonderem Masse zu. Je komplexer die angewendeten Verfahren der maschinellen Sprachdatenverarbeitung werden, desto tiefer, d.h. präziser werden Texte maschinell erschließbar. Hierdurch werden sich v.a. die "Referenz-Dokumentation" (d.h. die Erschließung textueller Dokumente über Deskriptoren oder Stichwörter) und "Fakten-Dokumentation" (hier: die Erschließung von Fakten aus Texten dergestalt, dass seitens des Systems ohne Heranziehung des originalen Inputs Problemlösungen zu einer Suchanfrage gegeben werden) zunehmend einander nähern (vgl. dazu auch KUHLEN 1979, S. 15 ff).

Beide Entwicklungen - die zunehmende Computerisierung wie auch die Vertiefung der Erschließung selbst - sind also als ein *Prozess* anzusehen; Differenzierungen, wie sie DIN 31623 vornimmt, stellen allenfalls Momentaufnahmen dar.

Eine Standortbestimmung der Automatischen Indexierung kann unter den verschiedensten Kriterien erfolgen. Für den vorliegenden Fall wird der *Bezug zur praktischen Anwendung* in den Vordergrund gestellt. Es geht dabei weniger um einen Vergleich zu intellektuell-manuellen Verfahren - so interessant dies für die Zukunft auch sein mag. Aus einer Reihe von Gründen werden

zudem *mathematisch-statistische Verfahren* ausgeklammert. Einerseits ist dies, v.a. die Probleme der Gewichtung von Deskriptoren und die Verknüpfung mit Cluster- und Klassifikationsverfahren, zentraler Gegenstand weiterer Beiträge dieser Jahrestagung (vgl. z.B. die Beiträge von FUHR, PANYR, AMBROSI/LAUWERTH und DEGENS). Zum Anderen setzen derartige Verfahren gerade voraus, dass der "Input", d.h. im vorliegenden Falle das zugrundeliegende Sprach-Datenmaterial, möglichst bereits um "oberflächige" Erscheinungen, bereinigt, d.h. tiefer aufbereitet ist.

Insofern wird die Betrachtung eingegrenzt auf *1inguistisch orientierte* Verfahren der automatischen bzw. computergestützten Texterschließung. Der Aspekt, unter dem dies versucht wird, ist weniger historisch als sprachanalytisch-methodisch orientiert. Die Verfahren, die exemplarisch vorgestellt (oder besser: angeführt) werden, erklären sich allerdings naturgemäß - zumindest in gewisser Weise - aus historischen. Gegebenheiten. So war die sog. Hardware (z.B. Computergeräte und Speicherkapazitäten) in den vergangenen zwei Jahrzehnten relativ teuer, die Rechengeschwindigkeit - im Vergleich zu heutigen Möglichkeiten - um den Faktor 1.000 oder 10.000 langsamer. Die sprachwissenschaftlichen Erkenntnisse und Methoden haben sich v.a.. in den 60-er und 70-er Jahren zunehmend "formalisiert", d.h. sprachliche Regeln bzw. Regularitäten wurden leichter an die Technologie und Logik der derzeitigen Computergeneration anpassbar. Aber auch wenn man dies alles in Rechnung stellt, ergibt sich dennoch eine zu große Diskrepanz zwischen der Methoden-. und Modellentwicklung in theoretischer Forschung bzw. Laboranwendung und der Wirklichkeit, d.h. der Anwendung automatisierter und computergestützter Verfahren zur Indexierung in der Dokumentationspraxis, als dass sie *allein* aus den angegebenen historischen Begrenzungen zu erklären wäre.

Das eigentliche Problem dürfte im *Transfer* der erarbeiteten Modelle in die praktische Anwendung liegen. Diese Transferproblematik hat nach aller Erkenntnis eine Reihe von Ursachen, die hier - wenn auch unvollständig - einmal mit Blick auf die Automatische Indexierung genannt werden sollen:

- Festhalten an "bewährten" Systemen: Wenn bereits intellektuelle Indexierungsverfahren (erfolgreich) zum Einsatz kommen (evtl. auch große Datenmengen damit erschlossen sind), so wird ein Wechsel zu neuer Technologie keine Priorität haben.

- Angst um Arbeitsplätze: In gewisser Weise muss der Bibliothekar oder Dokumentar um seinen Arbeitsplatz fürchten, wenn Maschinen seine Tätigkeit (weitgehend) übernehmen können.

- Unwissenheit bzw. unzureichendes Wissen um die Möglichkeiten (und Grenzen) neuer Technologie: Im Bereich des "kommerziellen" Information Retrieval sind echte Experten für computerlinguistische oder informationslinguistische Fragestellungen kaum anzutreffen; es dominieren (bislang) die EDV-Experten und Kaufleute. Eine negative Rolle spielen in diesem Zusammenhang die Medien. So bleiben meist nur anspruchsvolle (und zudem praktisch bislang kaum eingelöste) Verfahren, z.B. zur Maschinellen Sprachübersetzung oder zum "intelligenten" Verstehen geschriebener oder gesprochener Sprache, haften, dabei verknüpft mit Assoziationen wie "Spielereien" oder "Elfenbeinturm-Forschung".

- Scheu vor größeren Investitionen: Sprachverarbeitung mit dem Computer setzt auch (und gerade) unter der Bedingung, dass solide und ausbaufähige Verfahren zugrundegelegt werden, relativ große Investitionen sowohl in die Entwicklung der Verfahren selbst als auch bezüglich des Aufbaus des Datenmaterials (z.B. für die Fertigung maschineller ein- und mehrsprachiger Lexika) voraus. Kommerzielle Unternehmen, die letztlich zumindest auf Amortisation ihrer eingesetzten Mittel ausgerichtet sein müssen, haben bislang nicht die *hierzu nötige* Investionsbereitschaft gezeigt, im Gegenteil: die am Markt verfügbare Software im Bereich des Information Retrieval krankt allgemein an Ad-hoc-Lösungen und Provisorien, die dann auch noch teilweise in der "Überzeugung" entstehen, dass weitergehende Verfahren unnötig, nicht machbar oder zu teuer seien.

Im vorliegenden Zusammenhang sollen v.a. *vor dem Hintergrund der praktischen Anwendung die verwendeten Methoden und Verfahren zu einer computergestützten (bzw. automatischen) Indexierung auf der Ebene des Referenz-Retrieval zu textuellen Dokumenten* betrachtet werden. Diese Eingrenzung geschieht in der Annahme, dass dieser Bereich bei einem erfolgreichen Transfer des Know-how über die gegenwärtigen Möglichkeiten in die Praxis am ehesten die anstehenden Probleme zu lösen imstande sein wird.

Während man in der Fachinformation und -dokumentation für die wichtigsten Bereiche (z.B. Medizin, Chemie) wenigstens partiell über intellektuelle Verfahren der Klassifikation und Indexierung verfügt, die sehr nützliche Dienste bei der Informationsvermittlung leisten - allerdings auch speziell geschulte Indexierer und meist auch entsprechend "angepasste" Rechercheure erfordern, zeichnet es sich mit der Entwicklung von Büro- und Publikumskommunikationssystemen ab, dass hierzu intellektuelle Vermittlungsinstanzen nicht mehr ausreichen oder verfügbar sein werden.

Ein Beispiel dafür ist Bildschirmtext (BTX): Schon die Feldversuche in Berlin und Düsseldorf haben gezeigt, dass - v.a. mit Bezug auf den Abruf aktueller Informationen - das Suchbaumprinzip sowohl aus Gründen der Komplexität als auch aus Aktualitätsgründen hierfür unzureichend ist. Ein weiterer, in dieser Hinsicht völlig brachliegender Bereich ist die Bürokommunikation, insbesondere das Archivieren und Wiederfinden textueller Daten (Briefe, Notizen, Berichte, Mitteilungen, Protokolle usf.). Es dürfte allgemein einleuchten, dass in diesem Bereich eine intellektuelle "Indexierung" i.S. der für die Fachinformation entwickelten Verfahren und Modelle nicht möglich sein wird. Der für diesen Bereich vorhergesagte Markt wird es zudem erlauben, gegenüber den heute verfügbaren Verfahren verbesserte Systeme zu entwickeln. Diese müssen sich notwendig an den Problemen der natürlichen Sprache orientieren, auch um *systematisch* pflegbar zu werden, zudem mit dem im Bürobereich nicht unwichtigen Nebeneffekt, Fragen der Automatischen Silbentrennung und Rechtschreibhilfe bzw. -korrektur mitzubehandeln.

2. Stufen einer maschinellen Sprachdatenverarbeitung zur Indexierung

Die Aufgabenstellung der Automatischen Indexierung - in dem oben eingeschränkten Sinne - besteht einmal darin, ein System zu entwickeln, das auf weitestgehend *beliebige* natürlichsprachige Texte anwendbar ist. Dies betrifft zunächst die strukturelle Ausformulierung der Texte. Kurze wie lange Sätze, kurze und längere Wörter sollten bearbeitet werden können; es sollte (letztlich) auch keine Rolle spielen, ob ein Satz grammatikalisch ausformuliert ist (also z.B. ein Verb enthält) oder ob der Text frei von Rechtschreibfehlern ist. Derartige Anforderungen an die

Robustheit bzw. Flexibilität eines Systems führen zu folgenden trivialen Konsequenzen: Entweder sinkt bei sprachlich-technisch einfachen Lösungen die Qualität der Ergebnisse oder es werden inhaltlich wie technisch aufwendigere und damit kostspieligere Lösungen erforderlich.

2.1 Zeichenorientierte Indexierung

In der "traditionellen" Automatischen Indexierung vergleichbarer Zielsetzung wurden bislang weitgehend einfache Lösungen angestrebt. Am Markt (d.h. kommerziell) verfügbare Information-Retrieval-Systeme wie STAIRS (in seiner einfachen Form) oder DIRS/GRIPS sind hier beispielhaft zu nennen. Nach ihrer Verfahrensweise können sie bezüglich der Indexierungskomponente als *zeichenorientierte Systeme* betrachtet werden: Eine *Bearbeitungseinheit* ist dabei eine beliebige Kette aus Buchstabenzeichen, begrenzt von einem Zwischenraum oder Satz- bzw. Textzeichen. Operationen über den *Wortformen* (anhand einer Liste) ordnen diese zunächst zwei Mengen zu: der Menge der sog. *ST0P-Wörter* oder der Menge der einen Text (im Information Retrieval spricht man von einem "Text-Dokument' oder kurz einem "Dokument") charakterisierenden *Stichwörter*. Über den Stichwörtern operieren ggf. sog. *Stringverarbeitungs*funktionen, um diese Wortformen-Stichwörter auf Mengen gleicher Teilwortketten abzubilden (sog. Trunkierung). Dieses rein technische (d.h. *nichtsprachliche*) Verfahren soll der (Teil-)Schwierigkeit begegnen, dass bedeutungsgleiche oder -verwandte Wortformen eines Textes eine unterschiedliche Zeichenkette aufweisen (bekanntester Fall ist die sog. *Endungs*flexion, hinzu kommen Wort*ab1eitungen* und Wort*zusammensetzungen*).

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

2.2 Wortorientierte Indexierung

Die Konsequenz dieser letztlich unbefriedigenden technischen Lösungsansätze ist es, Modelle und Verfahrensweisen zu entwickeln, die Zeichenketten weitestgehend gleichen "Inhalts" (v.a. im Bereich der Flexion/Derivation) automatisch einander zuordnen. Hierzu lässt sich eine Reihe von Alternativen denken. Sie orientieren sich alle mehr oder minder an den Regeln zur *Woh1geformtheit* einer Zeichenkette in natürlichsprachigen Texten. Aufgrund der Erkenntnis, dass diese Wohlgeformtheit *nicht zuverlässig* über bedeutungs*unabhängige* Regeln zur Kombination von Buchstaben bei der Wortbildung möglich ist (auch wenn z.B. nicht *jede beliebige* Zeichenkette aus Buchstaben des Alphabets "physiologisch" sprechbar ist), und der Erfahrung, dass wegen der Möglichkeit der Integration fremdsprachiger (Lehn-)Wörter nicht *ein* Regelsystem alleine zugrunde gelegt werden kann, dass zudem die historisch gewachsene natürliche Sprache

viele Relikte heute nicht mehr regelhafter Wortbildungen aufweist, wird den diesbezüglichen Verfahren mehr oder minder ein *Vokabu1ar* (entweder in Form von ad-hoc entwickelten Ausnahmelisten oder systematisch in Form von Lexika) zugrunde gelegt.

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

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

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

Zur Lösung des besonderen Problems der *Identifikation von Mehrwortbegriffen* gibt es in nahezu allen kommerziellen IR-Systemen Hilfslösungen. Während boole-sche Verknüpfungen (d.h. mengenlogische Operationen wie UND, ODER, UND NICHT, z.B. HAUS UND VERKAUF, VERKAUF ODER VERLEIH; VERLEIH UND NICHT LEASING) i.a. dazu dienen, Dokumente zu identifizieren, die (irgendwie) Stich- oder Schlagwörter in der gewünschten Verknüpfung aufweisen, lässt sich vielfach die "Nähe" dieser Wörter zueinander (z.B. im gleichen Satz, unmittelbar nebeneinander) formal sehr einfach notieren und bei der Suche entsprechend ausnut-

zen. Es handelt sich also um eine ähnliche *technische* (und nicht sinnbezogene) Funktion wie die Trunkierung, wobei ausnutzt wird, dass *mehrwortige* Begriffe (z.B. JURISTISCHE PERSON, TREIBER IN DREIZUSTANDSLOGIK, METHODIK DES DEUTSCHUNTERRICHTS) auch physisch relativ "nahe" nebeneinander im Text vorkommen. Ähnlich der Trunkierung bedarf es beim Retrievalvorgang in derartig ausgerichteten Systemen entsprechender "intellektueller" Überlegungen und deren Umsetzung in Retrievalanweisungen, um Mehrwortbegriffe zu identifizieren.

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

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

2.3 Satz- und textorientierte Indexierung

Die Lösung der angesprochenen Probleme kann letztlich nur durch verbesserte Lexika und wortübergreifende (d.h. zumindest *satz*-orientierte) Verfahren erfolgen. Diese müssen v.a. zum Ziel haben,

- Wortformen auf Grundformen zurückzuführen, (z.B. HAUS, HAUSES, HÄUSER -- HAUS; WÄHLTE, GEWÄHLT -- WÄHLEN);

- Komposita (v.a. die im Deutschen so häufigen Augenblickskomposita ihren phraseologischen Synonymen zuzuordnen (z.B. BUNDESKANZLERWAHL - WAHL DES BUNDESKANZLERS);

- Wortableitungen miteinander in Beziehung zu setzen (z.B. WAHL - WÄHLEN - WÄHLER ...);

- Wortzusammensetzungen in sinnvolle (recherchierbare) Segmente zu zerlegen und entsprechend zu relationieren (z.B. BUNDESKANZLER - KANZLER);

- semantisch mehrdeutige Wörter zu vereindeutigen (z.B. ANLAGE, BANK, GUT, ...).

Exemplarisch soll hierzu ein System vorgestellt werden, das an der Universität des Saarlandes inzwischen im Rahmen langjähriger Forschungen entwickelt wurde und für deutschsprachige Texte entsprechende Funktionen bereithält. Es trägt den Namen CTX (für Computergestützte TeXterschließung). Dabei wurde das Ziel verfolgt, für die hier angesprochenen Problemkreise sprachlich motivierte und sprachbezogene (d.h. stärker "linguistische") Lösungen im Modell fachsprachenbezogen zu entwickeln und labormäßig zu erproben (Anwendung im Bereich Datenschutzrecht). Neben der systematischen Einführung linguistischer Lösungsansätze (z.B. zur *vollständigen* Behandlung der Flexionsmorphologie) sollte - auch dies zur Überwindung bestehender Ansätze - das System (z.B. im lexikalischen Bereich) offen sein für spätere Stufen, z.B. für eine maschinelle Sprachübersetzung.

Als Basis des Verfahrens wurde zu diesem Zweck das in grundlagenorientierter Entwicklung befindliche "Saarbrücker Übersetzungssystem" (SUSY), insbesondere in den auf die Analyse der deutschen Sprache bezogenen lexikalischen wie algorithmischen Teilen, herangezogen. Teile des SUSY-Systems (z.B. auch dort entwickelte Lexika) sind somit Teile des CTX-Systems, dessen zusätzliche Software den *sprachanalytischen Tei1* des SUSY-Systems gleichsam umrahmt (Inputvorbereitung,-Outputverarbeitung). Im lexikalischen Bereich wurde allerdings nicht nur eine (meist textbezogene, z.T. auch systematische) Erweiterung des Lexikoninventars im Rahmen der CTX-Entwicklung durchgeführt, sondern auch wesentliche Teile des Lexikon-Systems (v.a. die Entwicklung der Derivationslexika und des Thesaurus-Systems) neu konzipiert und materiell ausgefüllt. Das auf diese Weise entstandene Gesamtsystem "Computergestützte Textanalyse" (CTX) erbringt zum Abschluss der 2. Projektphase folgende *wesentliche Funktionen*:

- Flexionsformen werden *systematisch* auf eine sie repräsentierende *Grundform* zurückgeführt. (Dabei werden auch Um- und Ablaute, Infigierungen, abgetrennte Verbzusätze behandelt);

- Wortableitungen werden *systematisch* einander zugeordnet. (Dies geschieht durch *Relationierung* der Wörter, so dass dem Benutzer die Möglichkeit bleibt, diese Funktion einzubringen oder wegzulassen);

- Wortzusammensetzungen werden (unter intellektueller Kontrolle) lexikalisch (durch entsprechende Relationen) mit sinntragenden Teilwörtern verknüpft;

- Mehrwortbegriffe werden (auf der Grundlage bestimmter syntaktischer Strukturen) identifiziert.

Diese sowohl den RECALL erhöhenden als auch die PRECISION (bei Bedarf) steigernden Funktionen werden in jedem Anwendungsfall von CTX benutzt (Morphosyntaktisches System CTX-

I). Bei bestimmten Anwendungen - dies ist v.a. bei Integration eines systematischen Thesaurus erforderlich - wird ein weiterer Systemteil zur semantischen Disambiguierung benötigt, der den morphosyntaktischen Teil voraussetzt und integriert (semantisches System CTX-II). Damit sind für alle zuvor angesprochenen Problemfälle der Freitext-Indexierung entsprechende Funktionen entwickelt:

- Ermittlung von Grundformen und Aufbau von Wortableitungsrelationen bzw. Teilwortrelationen, statt Wortformen/Trunkierung

- Ermittlung natürlichsprachiger Mehrwortbegriffe anstelle komplizierter Abstandsfunktionen

- semantische Vereindeutigung mehrdeutiger Wörter statt boole-sche UND-Verknüpfung beim Retrieval (CTX-II).

Das Verfahren befindet sich gegenwärtig im Rahmen einer Transferphase in mehren praktischen Tests, u.a. mit Daten des Deutschen Patentamts (Bearbeitung von ca. 20.000 Offenlegungsschriften) und des Fachinformationszentrums FIZ Werkstoffe sowie des Wissenschaftszentrums Berlin. Erste Ergebnisse wie auch ähnliche Verfahren, wie z.B. das System INDEX2, das gegenwärtig in der DDR entwickelt wird (vgl. GRAICHEN 1981), zeigen, dass dieser Ansatz eine für die Praxis brauchbare Entwicklungsstufe zu sein verspricht.

Mit der linguistisch und sprachsystematisch ausgerichteten Bearbeitung der Flexionsmorphologie, der Derivations- und Kompositionsproblematik sowie der sprachbezogenen Erkennung von Mehrwortbegriffen ist von der wortbezogenen zur kontextuellen (zumindest phraseologischen) Sprachdatenverarbeitung übergegangen worden.

Im Bereich der semantischen Disambiguierung wird dabei ggf. nicht nur enzyklopädisches Wissen (z.B. kondensiert in einem Thesaurus), sondern auch der satzübergreifende Kontext von Bedeutung. Die dadurch angestrebte höhere Qualität (aber auch: die größere "Natürlichkeit" und Bequemlichkeit) beim Retrieval hat andererseits ihren Preis: mehr Aufwand in der System- und Lexikonpflege, größerer Aufwand an Rechenzeit und weiterer Speicherplatzbedarf. Dies kann an dieser Stelle nicht im Detail begründet und behandelt werden; manches "Negativum" (z.B. bezüglich der Rechenzeit- und Kodieraufwand) davon ist zudem gegenwärtig bedingt durch die vorgegebenen Labor- und Forschungssituationen. In der Diskussion mit Anwendern und Experten im Bereich des (dokumentorientierten) Information Retrieval spielen derartige Fragen allerdings eine wichtige Rolle.

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

schon Anfang der 60-er Jahre begründet worden); das lexikalische Wissen ist ebenfalls weitestgehend in traditionellen Lexika gespeichert. Zwei Gründe könnten insbesondere maßgebend gewesen sein für eine derartig späte Entwicklung: Einerseits ist jeder Laie - oberflächlich betrachtet - ein Experte in Sachen "Sprache". Ob er nun bewusst oder unbewusst die Regeln seiner Muttersprache beherrscht, so erscheinen sie ihm doch zu komplex und heterogen, als dass er sich von einem Computer hierzu der intellektuellen Leistung vergleichbare Ergebnisse und Funktionen vorstellen könnte. Die inzwischen kommerziell verfügbar gewordenen Ad-hoc-Lösungen auf Zeichen- und Wortebene bestätigen eher diese Vorstellungen als sie zu falsifizieren.

Auf der anderen Seite - und dies setzt sich bis in die Gegenwart fort - wurden linguistische Modelle in erster Linie - wenn überhaupt - zu Forschungszwecken in Computerprogramme umgesetzt. Im Vordergrund stand die Simulation von Sprachanalyse und -verstehen. Extreme Beispiele sind die Modelle Künstlicher Intelligenz, die bislang allenfalls auf kleine sprachliche wie physische "Welten" bezogen waren. Diese wissenschaftlich und insbesondere sprachwissenschaftlich sehr wohl begründeten Ansätze und Verfahren sind für die Lösung praktischer Probleme bislang weitestgehend ohne Wert geblieben. Eine Reihe weiterer Verfahren, die z.T. mit erheblichen Mitteln der öffentlichen Hand gefördert waren (z.B. das LIMAS-Verfahren von Hoppe oder das PLIDIS-System des Instituts für Deutsche Sprache) blieben entweder weitgehend in theoretischen (computerlinguistischen) Grundsatzfragen stecken oder zu anwendungsfern. Um es einmal mit der Herstellung von traditionellen Wörterbüchern zu vergleichen: Wenn man ein Wörterbuch für den Schulgebrauch machen möchte, kann man sich nicht auf die Lösung der Frage der Bedeutungsdifferenzierung konzentrieren und dies an einem Beispiel (etwa dem Wort "Liebe") exemplarisch erproben.

Umgekehrt reicht es allerdings z.B. für die Verwendung im Fremdsprachenunterricht nicht aus, den wesentlichen Wortschatz einer Sprache alphabetisch geordnet aufzulisten, vielmehr gehören Angaben zur Flexion, zu Wortfamilien (Komposita, Derivationen), zum syntaktischen Gebrauch, zur Betonung und Aussprache, evtl. zur Etymologie sowie Bedeutungserklärungen und Merkmale zu fachsprachlichen und stilistischen Besonderheiten dazu. Auf den Bereich der Automatischen Indexierung übertragen bedeutet dies, die gegebenen und bewährten linguistischen Erkenntnisse stärker zu nutzen und sich nicht in Modellen zu verlieren, zugleich den praktischen Bedarf (etwa im Bereich des Referenz-Retrieval) bzw. der Freitextanalyse gerade anhand der Ad-hoc-Lösungen zu erkennen und daraus ein Konzept der schrittweisen Einführung komplexerer Verfahren abzuleiten.

3. Perspektiven der weiteren Entwicklung

Wenn man - wie hier - bei der Automatischen Indexierung für ein stufiges Vorgehen plädiert, so muss man sich bewusst sein, dass jede Stufe (die in sich, nebenbei bemerkt, wiederum Variationen aufweisen kann, also eher als abstraktes Konzept zu sehen ist) Unvollkommenheiten und Teillösungen in sich birgt, die durch tiefere Erkenntnisse bzw. Funktionen überholt werden können. Die stufige Vorgehensweise hat jedoch den Vorteil, dass *jewei1s* praktikable, d.h. für verschiedene Anwendungen (und letztlich immer mehr Fragestellungen) nützliche (Zwischen-)Ergebnisse erreicht werden. Im Gegensatz zu "reinen" Ad-hoc-Ansätzen (wie sie im Bereich des kommerziellen Information Retrieval häufig zu beobachten sind) sollte jedoch die stufige Vorgehensweise so weit *1inguistisch motiviert* erfolgen, dass die erreichten Teillösungen *integraler Bestandteil weitergehender Konzepte* werden können. Dies unterscheidet z.B. die Vorgehens-

weise bei CTX von derjenigen anderer Verfahren (z.B. sei hier auf die Key-Phase-Technik bei SEELBACH 1975 oder die Methoden des Partiellen Parsing in ROSTEK 1979 verwiesen).

Die systematische Entwicklung höherwertiger linguistisch motivierter Verfahren zur Automatischen Indexierung wird in Zukunft zunehmend aufgrund zweier entscheidender Prozesse Unterstützung erfahren: Es handelt sich einmal um die bekannte Miniaturisierung der Computertechnik unter Ausweitung der Speicher- und Zugriffstechniken zu größeren Datenbeständen. So ist es heute bereits praktizierbar, maschinelle Wörterbücher als Instrument der Rechtschreibhilfe und Silbentrennung auf Textverarbeitungssystemen einzusetzen (man vgl. die Entwicklungen im kommerziellen Bereich, z.B. des IBM-Schreibsystems, bei ALPHATEXT oder auch bei NIXDORF). Für diesen Bereich der Bürokommunikation werden zwar zunächst weniger komplexe Teillösungen entstehen mit der Ausweitung auf Textarchivierung und -retrieval in Büro und Verwaltung werden zunehmend aber auch höherwertige (d.h. über die wortorientierte Verarbeitung hinausgehende) Verfahren einbezogen werden.

Der zweite - anhaltende - Prozess, die stetige Kostensenkung im Bereich der Computer-Hardware, verbunden mit dem wachsenden Kostenanstieg im Personalbereich, lässt (verbesserte) Verfahren einer Automatischen Indexierung zunehmend attraktiver werden in den Fällen, in denen heute noch die intellektuelle Texterschließung dominiert.

Als nächste "Stufen" könnten sich *multilinguale Indexierungsverfahren* entwickeln, die zumindest auf *Begriffsebene* Mehrsprachigkeit (beim Retrieval) zulassen. Im einfachsten Falle kann dies über Synonymie-Relationen im Thesaurus geschehen, wobei jedoch das Problem der einzelsprachlichen Mehrdeutigkeit zu lösen ist. Eine weitere Entwicklung könnte die unmittelbare Verbindung mit maschinellen Übersetzungssystemen bedeuten. Hierbei wird allerdings bereits der "klassische" Indexierungsbereich verlassen, eine Erkenntnis, die sich jedoch bereits bei der satz- und textorientierten Indexierung abzeichnet. Indexierung stellt in diesem Sinne einen Spezialfall der "Übersetzung" dar, bei dem ähnliche Prozesse der Analyse, des (sprachlichen) Transfer und der Sprachsynthese (d.i. der Deskriptor-Erstellung) auftreten.

In diesem Zusammenhang sind natürlich im Grunde mögliche Entwicklungen anzusprechen, die *weitergehende* Inhaltsanalysen i.S. einer "Verdichtung" der textuellen Information durch tiefere und erweiterte linguistische und/oder statistische Verfahren betreffen. Angesichts der Vielfalt der möglichen Relevanz(en) eines Textes (Aufsatzes, Briefes, Berichts, Protokolls) bezüglich bestimmter Suchanfragen und Problemlösungen wird abzuwarten sein, inwieweit derartige Forschungen - die bezüglich des "linguistischen" Teils noch eher zur Grundlagenforschung zu rechnen sind - einmal praxisrelevant sein werden. Angesichts der zunehmenden "Computerisierung" in Wirtschaft, Wissenschaft und Verwaltung wird der Bedarf nach derartigen verfeinerten Verfahren sicherlich in naher Zukunft jedoch gewaltig anwachsen.

Quellen:

Ambrosi, K.; Lauwerth, W. (1983): Ein Klassifikationsverfahren bei qualitativen Merkmalen. In diesem Band.

CTX - Ein Verfahren zur Computergestützten Texterschließung. BMFT-Forschungsbericht (im Erscheinen).

Degens, P.O. (1983): Hierarchische Clusteranalyse: Eigenschaften und Berechenbarkeit. In diesem Band.

DIN 31623: Indexierung zur inhaltlichen Erschließung von Dokumenten (Entwurf). Deutsches Institut für Normung e.V. (DIN). Berlin 1978.

Fuhr, N. (1983): Klassifikationsverfahren bei der Automatischen Indexierung. In diesem Band.

Graichen, D. (1981): Thesaurusunabhängiges Indexieren medizinischer Befunde mit "INDEX2". In: Informatik 28 (1981) 4, S.30-35.

Kuhlen, K. (1977): Experimentelle Morphologie in der Informationswissenschaft. München.

Kuhlen, R. (1979) Hrsg.: Datenbasen, Datenbanken, Netzwerke. Praxis des Information Retrieval, 1: Aufbau von Datenbasen. München.

Panyr, J. (1983): Automatische Indexierung und Klassifikation. In diesem Band.

Rostek, L. (1979): Methoden des partiellen Parsing für die automatische Indexierung - Syntaxgraphen zur Analyse von Sprachmustern. In: Kuhlen 1979, S. 251-282.

Salton, G. (1971): The SMART Retrieval System. Englewood Cliffs; New Jersey.

Salton, G. (1981): A Blueprint for Automatic Indexing. In: SIGIR Forum 16 (1981) 2, S. 22-38.

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

Zimmermann, H. (1979): Ansätze einer realistischen automatischen Indexierung unter Verwendung linguistischer Verfahren. In: Kuhlen 1979, S. 311-338.

Diskussion:

Ohly fragt an, warum nicht über die statistischen Verfahren gesprochen worden sei. Zimmermann erläutert, dass er sie bewusst ausgeklammert habe, da ja darüber im nächsten Vortrag Herr Fuhr sprechen werde. Was jedoch seine Bewertung dieser Verfahren betrifft, so halte er sie für eine gute Ergänzung, aber man komme wohl nicht allein damit aus. Sie seien sinnvoll für eine Erstorientierung.
Endres bezieht sich auf eine Bemerkung in Herrn Fugmanns einleitenden Worten über die wenig befriedigende intellektuelle Indexierung bei großen Datenbanken und bittet um Belege für derartige Aussagen in der Fachpresse. Fugmann verspricht dies. Gleichzeitig weist er auf drei Sachverhalte hin: 1. Wo gibt es eine intellektuelle Indexierung, die sich an vorgegebenen semantischen Kategorien orientiert? Es gibt sie in der Praxis äußerst selten. 2. Wo findet man in den verwendeten Klassifikationsschemata klar herausgearbeitete Einteilungsgründe? Äußerst selten. Meistens sind die Klassifikationstafeln - außer denen der indischen Schule - ein Sammelsurium von irgendwie zusammenhängenden Begriffen, bei denen jedoch nicht klar herausgearbeitet wurde, wie sie zusammenhängen. 3. Wo findet man Vokabu1arien, nach denen sich der Indexer und der Fragesteller richten können, bei denen vorgeschrieben ist, nicht irgendeinen, sondern den bestpassenden Terminus aus dem Vokabular zu wählen? - Auch dies setzt wiederum einen hohen Ordnungsgrad in diesen Vokabularen voraus. Hier liegen also nur 3 Ansatzpunkte für eine drastische Verbesserung vor, die die Zuver1ässigkeit einer Indexierung wesentlich steigern könnten. Zimmermann wollte wissen, ob der Endbenutzer ein solches System beherrschen müsste oder ob es eines Vermittlers bedürfe? Fugmann erwidert, dass er keine Anzeichen dafür erkennen könne,

dass in Zukunft ein Informationsvermittler entbehrlich würde. Es gäbe eine große Zahl von Fragen, die auch ein Informationslaie beantworten könne, aber es gibt auch eine große Zahl von Problemen, bei denen der weniger Erfahrene hoffnungslos überfordert sei und die Antworten gerne Experten überlässt. Zimmermann: Können Sie sich vorstellen, dass der Experte auch ein Computer sein kann? Fugmann: Dazu möchte ich keine Prognosen geben, sondern hier zuerst prüfen und gegebenenfalls auch zurückweisen können.
Scheele weist darauf hin, dass man die automatische Indexierung ja auch verknüpfen könne mit einer Notationsgebung. Bereits vor 20 Jahren wurden in der Biologiedokumentation 275 000 Titel automatisch klassifiziert und damals lautete die Anforderung, man muss auch beim Retrieval die Möglichkeit haben, nach Oberbegriffen und nach den Verknüpfungen/Kombinationen von Oberbegriffen zu fragen und zwar auf der 3., von Herrn Zimmermann geschilderten Ebene, nämlich der Verknüpfung von Wörtern eines Satzes. Sein Verfahren leistet dies. Zimmermann glaubt unterscheiden zu müssen zwischen einer Klassifikation und einer Verknüpfung. "Wir verwenden die natürlichsprachigen Begriffe, soweit sie vorhanden sind, um solche Relationen auszudrücken." Man könne sich somit auch ein Informationssystem vorstellen, das über solche Relationen auch in der natürlichen Sprache etwas erreicht. Was jedoch Herr Fugmann bezüglich des Problems der Verursacher und der Rollen, usw. meine, das bedeute die nächste Ebene, eine noch feinere Beschreibung. Wieweit sich solche Verfahren automatisieren lassen, wird erst die Zukunft zeigen. Aber, wenn man etwas schon formalisieren könne, dann sei auch die Möglichkeit gegeben, etwas zu automatisieren.