Harald H. Zimmermann

LINGUISTISCHE VERFAHREN ZUR ARCHIVIERUNG UND ZUM WIEDERFINDEN UNSTRUKTURIERTER TEXTE

Saarbrücken, April 1983

Abstract

Die technologischen Entwicklungen der 80-er und 90-er Jahre werden den Bedarf an "intelligenten" Verfahren zur automatischen Texterschließung und -archivierung sprunghaft steigen lassen. Inzwischen existiert eine Reihe linguistischer Verfahren, die auch große Datenmengen, wie sie in der Büro- und Fachkommunikation auftreten, zu bewältigen helfen. Dabei ist eine wesentliche Zielsetzung, den Anwender von "technischen" Handhabungen, wie sie die herkömmlichen Informations-Retrieval-Systeme noch erfordern, schrittweise in Richtung auf einen eher natürlichsprachigen Zugang zu entlasten. Während in den nächsten Jahren "verstehensorientierte" Ansätze nur in ausgewählten Bereichen zum Einsatz kommen können, werden Verfahren auf morphologisch-syntaktischer Basis die bisherigen oberflächenorientierten Systeme zunehmend ersetzen. Derartige Verfahren werden ihren Markt nicht mehr allein in der Fachinformation, sondern zunehmend auch in der Bürokommunikation finden.

1. Einführung

Im Frühjahr 1977 fand in Gießen eine Frühjahrstagung des Fachbereichs medizinische Informatik der Gesellschaft für medizinische Dokumentation und Statistik (GMDS) mit dem Fachausschuss 14 der Gesellschaft für Informatik (GI) unter dem Thema "Klartextverarbeitung" statt /1/. Der Anlaß war typisch für die Situation in vielen Fachgebieten: Hier ist z.B. die schriftlich fixierte Krankengeschichte Ausgangspunkt für "Beobachtungen und Interpretationen vieler an der Diagnostik und der Therapie beteiligter Personen" /2/. Das Führen einer Krankengeschichte erzeugt ein Dokument, "das die Krankenversorgung in der täglichen Routine und in der Forschung bezüglich Kommunikation, Rechtfertigung und Analyse unterstützt" /3/. Gelingt es nicht, Verfahren zu entwickeln, die aus derartigen frei formulierten oder auch in leicht vereinfachter natürlicher Sprache ausgedrückten Texten die notwendigen Fakten erschließen, so bleibt "nur die Alternative, dass die Daten von Menschen in eine stark formalisierte Sprache übersetzt werden müssen" /)4/ . Dies erscheint jedoch kostenaufwendig bzw. fehleranfällig und ist vielfach mangels geeignet geschulten Personals nicht zu leisten.

Dass Frei- oder Klartextverarbeitung /5/ nur einen Teilbereich der Informationsverarbeitung darstellt, braucht nicht weiter betont zu werden. Die medizinische Information und Kommunikation ist ein Beispiel dafür, dass auch der Auswertung von Analogdaten (Messungen, Graphiken) durch numerische und assoziative Prozesse (Mustererkennung) eine große Bedeutung zukommt. Wenn im folgenden derartige Komponenten ausgeklammert werden, so muss doch deutlich bleiben, dass letztlich nur eine integrierte Lösung, d. h. ein Konzept, das die Verarbeitung textueller, numerischer und bildlicher Information ermöglicht, den Anforderungen an ein Informationssystem gerecht werden kann.

2. Texterschließung und -retrieval in der Fachinformation

Der Prozess der Archivierung und des Wiederfindens von unstrukturierten Texten hat in der Fachinformation bereits eine große Tradition. Generationen von Spezialisten im Bibliotheks- und Dokumentationsbereich haben sich damit befasst. Angesichts der Fülle an spezifischem, in natürlichsprachlicher Form codiertem "Wissen" ist umgekehrt der Informationssuchende für nahezu jede Hilfe aufgeschlossen, die ihm bei der Suche nach Problemlösungen geboten wird.
Eine vermeintlich einfache Fragestellung setzt in der Praxis gegenwärtig eine Reihe von informationsstrategischen Umsetzungen voraus, ohne deren Realisierung die gewünschte textuelle Information nicht gefunden werden kann. Dies soll zunächst an einem "Beispiel" verdeutlicht werden. Die Fragestellung soll lauten: "Welche Erkenntnisse brachte der Bildschirmtext-Feldversuch in Düsseldorf bezüglich BTX im Hinblick auf den Bereich "Unterhaltung"?

"Idealerweise" stellt man eine derartige Frage an einen Experten, hier z.B., an einen Mitarbeiter im Team der wiss. Begleituntersuchung zu Bildschirmtext. Bei dem großen Interesse, das BTX gegenwärtig genießt, könnte die Beantwortung derartiger Fragen allerdings leicht mehr als einen Experten/Tag erfordern.

Man kann sich daneben an eine Bibliothek wenden mit dem vagen Verdacht, dass bereits eine Publikation zu dem Thema vorliegt. Bei einer fachspezifisch ausgerichteten Infrastruktur, wie es die sog. Fachinformationszentren im Grunde darstellen, ist eine entsprechend umformulierte Frage auch an ein Experten-Vermittlungssystem zu richten. Für die vorliegende Fragestellung kommt z.B. das Informationszentrum der Gesellschaft für Information und Dokumentation (GID) in Betracht. Dort sind inzwischen Informationsbanken verfügbar, die Hinweise zu Publikationen in "verdichteter" Form (z.B. Titelangabe, Kurzfassung, Schlagwörter) enthalten. Voraussetzung ist allerdings, dass die Information auf dem Wege der "Verdichtung" auf Abstracts, Schlagwörter und/oder Kategorien nicht verlorengegangen ist. Die Schlagwortsuche, ggf. kombiniert mit Begriffen, die möglicherweise im Titel einer Literaturangabe vertreten sind, lässt z.B. im vorliegenden Beispiel hoffen, dass einige sog. "Dokumente" zu den Begriffen "BTX", "Düsseldorf" und "Feldversuch" identifiziert werden, auch für den Fall, dass die zusätzliche Verknüpfung mit "Unterhaltung" keinen Treffer bringt. Der Informationsvermittler (evtl. auch der Informationssuchende) muss entsprechende Strategien anwenden, um die ggf. relevante Literatur zu exzerpieren.
Ein besonderer Komfort derzeitiger Systeme - etwa gegenüber einem Bibliothekskatalog - ist das Abstract, d. h. die kurze (u.U. auch kritisch-bewertende) Zusammenfassung einer wissenschaftlichen Arbeit. Nützlich immer dann, wenn daraus ersichtlich wird, dass der Primärtext wohl nicht zur Problemlösung beitragen kann. Bleibt allerdings die Möglichkeit bestehen, dass eine Arbeit für die Problemlösung relevant ist, muß üblicherweise ein Bestell- oder Ausleihvorgang eingeleitet werden, der im günstigen Falle (wie bei dem genannten Beispiel GID/IZ) recht kurze Wege hat, u. U. aber über mehrere Instanzen im Bestellverkehr der Bibliotheken führen kann.

Ein Wissenschaftler muß heute tief in die Gefilde der weltweit entstandenen Informationssysteme hinabsteigen und ihre verschlungenen Pfade gut kennen um schließlich, vielleicht auch als "Blinder" (Endbenutzer) von einem "Lahmen" (dem Informationsvermittler) geführt - wie durch ein Wunder zu den ersehnten Quellen der Erkenntnis und des Wissens zu gelangen.

3. Neuere Entwicklungen der automatischen Texterschließung

Es ist gegenwärtig eine noch offene Frage, ob Verfahren "höherer" Intelligenz, wie sie z.B. in den sog. "Frage-Antwort-Systemen" modellhaft erprobt werden, in diesem Jahrhundert noch in die breite Anwendung umgesetzt werden können. Derartige Verfahren lassen heute für "Modellwelten" schon faktenorientierte Abfragen zu; allerdings ist das Systemwissen noch wenig umfangreich, so dass eine konkrete, "verstehensorientierte" Beantwortung einer Anfrage (wie die im Beispiel gestellte) über ein System wohl noch Spekulation bleibt. Auch im Bereich der Abstracterstellung werden - sofern nötig - auf längere Sicht die intellektuellen Verfahren im Vordergrund stehen, soweit große bzw. heterogene Texte und Textsorten betroffen sind. Insofern wird dieser Bereich der "Künstlichen Intelligenz" bezüglich der Problemlösung im folgenden ausgeklammert.
Umgekehrt ist im Rahmen der intellektuellen Inhaltserschließung ein breites Feld von Verfahren vorhanden, die wenigstens teilweise in einem weitergehenden Ansatz zur automatischen bzw. computergestützten Texterschließung bereits heute erfolgreich berücksichtigt werden können. Eine "Mechanisierung" derartiger Verfahren bzw. eine auf dem sog. Referenzretrieval aufbauende automatisierte Texterschließung bietet den Vorteil, nicht nur in der Fachinformation Verwendung finden, sondern auch auf intellektuell nicht zu bewältigende Bereiche wie z.B. die Bürokommunikation angepasst werden zu können. Ein weiterer Vorteil ist, dass derartige Verfahren unmittelbar auf Primärdaten (d.h. dem Volltext) aufsetzen können.

Auch wenn einige Ansprüche, die aus den Anforderungen und Verfahrensweisen erwachsen, z.B. die Möglichkeit, die "Dokumentgröße" je nach Bedarf auf einen angemessenen "Kontext" (Buch, Kapitel, Absatz, Satz...) zu variieren, heute noch kaum von den (literaturbezogenen) IR-Systemen erfüllt werden, wird diese Entwicklung letztlich angesichts der wachsenden technischen Möglichkeiten der elektronisch-textuellen Kommunikation entscheidend an Boden gewinnen.

3.1 Textstruktur

Um die Frage des Einsatzes linguistischer Verfahren für Textarchivierung und -retrieval behandeln zu können, sollen im folgenden einige wesentliche Probleme dargestellt und exemplifiziert werden, die für das weitere Verständnis von Bedeutung sind.

Ein Text besteht - formal betrachtet - aus einer Folge von Wortformen, Satz- und Wortzeichen sowie Textmarkierungen wie z.B. Absatz- und Kapitelbegrenzern. Er ist in der Regel untergliedert in Kapitel und Absätze, die wiederum in Sätze gegliedert sein können.

Die äußerlich erkennbare Struktur muß nicht notwendig (wird aber in der Regel) konform gehen mit den Inhalten (d.h. den behandelten Themen) ; so können Absätze thematisch enger zusammengehören oder unterschiedliche Themen behandeln, ein verwendetes Wort kann für das Thema besonders relevant sein oder aber irrelevant usf. (Nahezu) Gleiche/identische Inhalte können in einem Wort oder einer Wortgruppe oder einem Satz ausgedrückt sein (z.B. BERLINREISE - REISE NACH BERLIN - "DIE REISE GEHT NACH BERLIN - EINE REISE, DIE BERLIN ZUM ZIEL HAT - X REIST NACH BERLIN). Daneben tritt gleichsam als Umkehrung der Ausdrucksvariation bei (nahezu) gleichem Inhalt in allen natürlichen Sprachen das Problem der Mehrdeutigkeit sprachlicher Teilstrukturen (Wörter, Phrasen, Sätze) auf. Wörter wie ANLAGE, GUT haben - lexikalisch betrachtet - mehrere Bedeutungen, die erst aus dem weiteren Kontext heraus oder aufgrund der thematischen Eingrenzung differenziert werden können. All dies sind

Fragen, deren Lösung bzw. Nicht-Lösung von Einfluss auf die Qualität und Leistungsfähigkeit eines textorientierten Archivierungs- und Retrievalsystems sein kann. Der Katalog derartiger Probleme kann sehr leicht um weitere, "höherwertige" Fragen erweitert werden. Hierzu rechnen wir das Problem der Textkondensation, d. h. der Zusammenfassung umfangreicher sprachlicher Aussagen auf das für wesentlich Gehaltene, das Problem der Sprachbarrieren, z.T., verbunden mit der Übersetzung eines Textes in eine andere natürliche Sprache, sowie das Problem der Wissensrepräsentation, d. h. der Einordnung des im Text Dargestellten in ein Wissensspeicher- und -verstehenssystem, wobei Veränderungen dieses (Welt-)Wissens entstehen und weitgehend von den natürlich-sprachlichen Ausdrücken abstrahiert wird.

Es ist ein Ziel der sprachwissenschaftlichen, d. h. linguistischen Betrachtung, die Sprache als ein System zu analysieren und zu begreifen, das dazu dient, (menschliches) kognitives Wissen zu kommunizieren, d. h. anderen Menschen zu vermitteln. Dass die (allgemeine) natürliche Sprache keineswegs ein perfektes Instrument hierfür ist, zeigt die Entwicklung hochspezialisierter Fachsprachen (z.B. in der Chemie, der Mathematik), zeigen aber auch die verschiedensten Fachterminologien. Ihr Ziel ist es im Grunde, den sprachlichen "Code" zu normieren, d. h. auch, den Code explizit zu machen, den (zwei) Sprachteilnehmer benutzen, wenn sie miteinander (fachliches) Wissen austauschen. Der unbewusst erlernte natürlichsprachige Code erscheint zu ambig, zu redundant, zu fehleranfällig. Teil und Aufgabe eines jeden Fachstudiums ist es somit (ohne dass man sich dessen vielleicht während des Studiums der Fakten so recht bewusst ist), eine neue "Sprache" (die Fachsprache) so zu lernen, dass man sich in ihr kompetent bewegen kann. Da umgekehrt der Ausschnitt der Fachwelt nicht immer präzise definiert werden kann, zusätzlich der Gegenstand (u.a. im Bereich der Forschung) nicht notwendig eindeutig präzisiert werden kann, ist auch eine Fachsprache (insbesondere ein Klassifikationssystem) zumindest offen, vielfach auch vage; eine Fachsprache, die in die natürliche Sprache eingebettet ist, unterscheidet sich in diesem Sinne nur graduell von dem natürlichen Standard-Sprachsystem.

Eine wesentliche Eigenschaft im Zusammenhang mit Sprache und sprachlichem Verstehen, die bisher auf der Suche nach einem eindeutigen Interpretationsmechanismus zu wenig beachtet wurde, im vorliegenden Falle jedoch besonders wichtig erscheint, ist das Phänomen des Zusammenhangs sprachlicher Vagheit und sprachlichen Verstehens. Es gibt eine Reihe von Tests, die zeigen, dass sprachliche Ausdrücke - z.B. bezogen auf Gegenstände - nicht anhand eindeutiger Merkmale (mengentheoretisch gesehen) klassenbildend sind (z.B. der Begriff "Stuhl" als Sitzgelegenheit). Es muß also (vielleicht analog oder identisch mit den Prozessen des Spracherwerbs) ein intellektuelles Verfahren existieren, das Aussagen auch dann zu "verstehen" erlaubt, wenn sie im Grunde - bei voller Anwendung eines präzisen sprachlichen Regelapparates - zu Nicht- oder Fehlinterpretationen führen würden. Ein sehr einfaches Beispiel ist das Tolerieren von Rechtschreibfehlern (ein Problem, das heute noch vielen computergestützten Systemen sehr zu schaffen macht), aber auch das Verstehen von "Gastarbeiterdeutsch" u.a.m.

3.2 Verfahren der Texterschließung

Vor diesem – skizzierten - Hintergrund soll nunmehr versucht werden, intellektuelle wie auch computergestützte bzw. maschinelle Verfahren von Texterschließung und -retrieval ansatzweise einzuordnen und zu bewerten.

(1) Intellektuelle Schlagwortvergabe (Indexierung)

Die intellektuell (d.h. durch den menschlichen Bearbeiter) vollzogene Abbildung von Texten bzw. Textwörtern auf Schlagwörter setzt i.d.R. einen Fachexperten voraus, der sich zudem an der Fachterminologie (z.B. in Form eines Schlagwortkataloges bzw. eines Thesaurus) orientiert. Die Auswahl der vergebenen Schlagwörter in Relation zur Textgröße führt ggf. zu einer entsprechenden Selektion von Termini (positiv gesehen: einer Gewichtung zur Vermeidung von Ballast, negativ gesehen: zu einer Informationsreduktion, d.h. zu einem Informationsverlust).

Das Retrieval (d.h. das Wiederfinden von Texten/Dokumenten) setzt wiederum einen Experten voraus. Ist die Terminologie des Systems Allgemeingut der Fachwelt, so kann dies ein "Endbenutzer" (d.h. der eine Problemlösung suchende Wissenschaftler) sein; ist sie (z.B. bei Vergabe von sog. "Vorzugsbenennungen") systemimmanent, so ist u.U. ein Systemexperte als Vermittler auch beim Retrievalvorgang erforderlich.

(2) Maschinelle Schlagwortvergabe (Indexierung)

Die herkömmlichen maschinellen Verfahren sind orientiert an den Zeichenketten eines Textwortes. Abgesehen von der Eliminierung weniger sog. STOP-Wörter (im Deutschen z.B. DER, UND, EIN ...) werden alle Textwörter in der äußeren Gestalt invertiert, d. h. in der jeweils vorgegebenen Wortform in einen Index (Register) übertragen mit Verweis auf den Text bzw. die Textstelle ihres Auftretens.
Es bleibt dem Informationssuchenden überlassen, geeignete Maßnahmen zu ergreifen, um anhand dieser Indexwörter möglichst viele relevante Texte wiederzufinden. Hierzu gehört zunächst die sog. Truncation (Trunkierung). Dabei handelt es sich u.a. um das Markieren von Stellen einer Wortform, ab der die auftretenden Zeichen vernachlässigt werden können (z.B. die Wortendungen bei "BILD": bei Markierungen durch ein Sonderzeichen, hier "$", also Bild, werden auch Wortformen wie BILDES, BILDERN usf. gefunden). Ein weiteres Mittel ist die Verwendung eines Wortabstandesmaßes, um mehr oder weniger "benachbarte" Wörter in den Retrievalprozess einzubeziehen. In beiden Fällen wird an die Kompetenz des Benutzers appelliert, aufgrund seiner "Grammatik" zumindest alle üblichen bzw. wahrscheinlichen "Paraphrasen" (d.h. die oberflächigen Ausdrucksvarianten bei "gleichem" Sinn) vorwegzunehmen und durch entsprechende Markierungen in den Suchauftrag an das System einzubeziehen.
Ähnliches gilt für den Ballast, der dadurch entsteht, dass Mehrdeutigkeiten nicht aufgelöst sind oder dass Wörter zwar im Text verwendet werden, thematisch aber ohne Belang sind usf. Auch hierfür wird die Kompetenz des Benutzers, die Spreu vom Weizen zu trennen, ausgenutzt. Ein Vorteil derartiger "unintelligenter" Verfahren ist die Einfachheit, besonders bezüglich der technischen Realisierung. Aber auch der Arbeits- und Pflegeaufwand ist gering: So ist z .B. keine Terminologieliste zu pflegen, kein Lexikon aufzubauen usf. Ein weiterer Ausbau, z.B. in Richtung auf einen maschinellen Thesaurus oder auf eine maschinelle Übersetzung, ist damit allerdings nicht möglich.

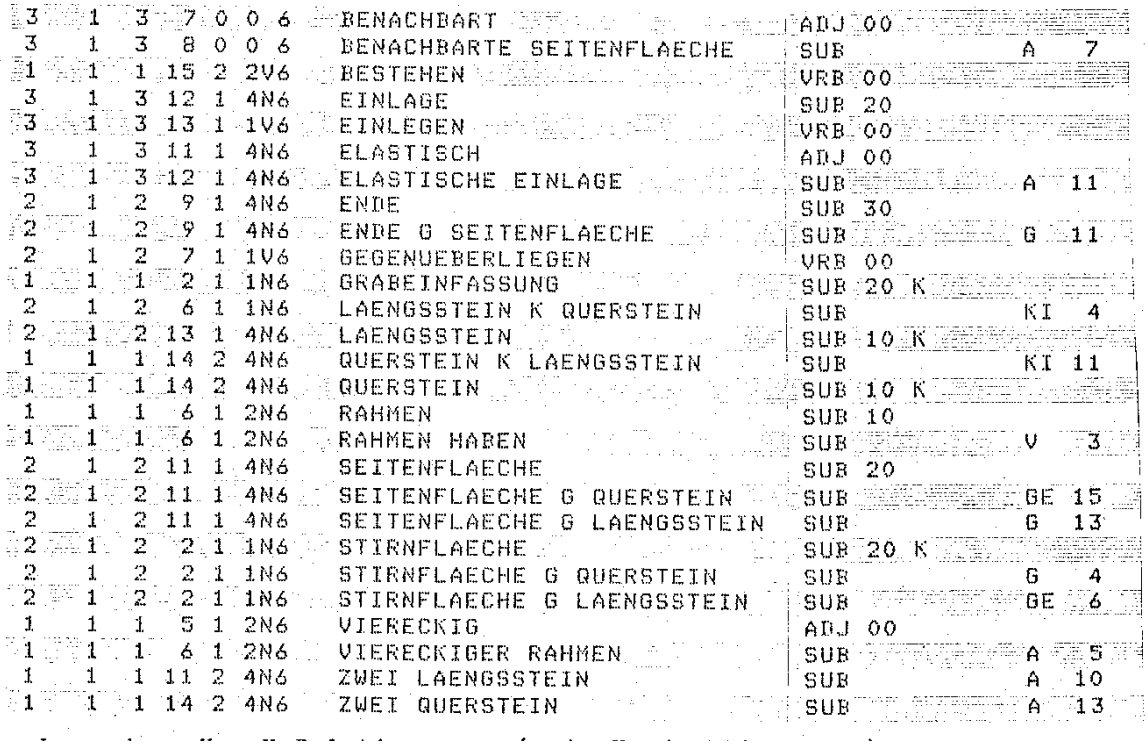
Eine "höhere" Stufe stellt die Integration mehr oder minder komplexer maschineller Lexika dar. Dies kann z.B. während der Phase der Indexierung geschehen (z.B. bei PASSAT /7/ oder CTX /8/), oder aber während des Retrievalvorgangs, wie z.B. bei der STAIRS-Variante TLS /9/.
Dabei soll zunächst die sog. morphologische Komponente berücksichtigt werden: Bei natürlichen Sprachen wie dem Deutschen tritt neben die Wortbeugung (d.h. Flexionsendungen mit Umlaut und Ablaut, z.B. HAUS, HAUSES, HÄUSER, GEHEN, GING, GEGANGEN) die Wortzusammensetzung (HAUSTÜR, TÜRSCHLOSS ...), bei einer Reihe von Sprachen ist die Zusammenführung von Wortableitungen (SCHÖN, SCHÖNHEIT; ABNEHMEN, ABNAHME,

ABNEHMBAR ...) im Hinblick auf eine "normierte" Abfrage von Bedeutung. Kann eine derartige Prozedur (wo auch immer) in ein System integriert werden, so wird der Benutzer von einer Reihe von Trivialitäten (z.B. bezüglich der Trunkierung) entlastet. Voraussetzung ist jedoch der Aufbau eines entsprechenden Wörterbuches und dessen Pfleger zudem sollte das Problem der sinnlosen Wortzerlegungen (z.B. PROZESSHANDLUNGEN nur in PROZESS und HANDLUNG, nicht aber in LUNGE und HAND) angemessen berücksichtigt werden. Derartige Verfahrenstechniken sind z.B. bei PASSAT und CTX in den Indexierungsprozess integriert. Dies bietet zusätzlich den Vorteil, dass Rechtschreibfehler (sofern sie nicht wieder sinnvolle Wörter ergeben), beim Aufbau des Informationsspeichers erkannt und korrigiert werden können.

Eine weitere Stufe stellen Verfahren dar, die automatisch mehrwortige Begriffe (unabhängig von der Flexion) erkennen oder Begriffe über syntakto-semantische Beziehungen verknüpfen können. Während die Grundformermittlung bzw. Kompositumzerlegung es erlaubt, die Trefferquote zu erhöhen (Verbesserung des sog. Recall), ermöglicht die Vergabe von Relationen zwischen Textwörtern aufgrund von im Text identifizierten Beziehungen eine größere Präzisierung und damit thematische Einschränkung beim Wiederfinden der Texte. Sie sollte sich in der frühzeitigen Vermeidung von Ballast niederschlagen (d.h. einer Erhöhung der sog. Precision). An die Stelle formaler Kriterien (wie z.B. des Wortabstandsmaßes) treten auf diese Weise Formulierungen, die dem natürlichsprachigen Ausdruck nahekommen. Ein Beispiel dafür ist wiederum die Verfahrensweise von CTX (Computergestützte Textanalyse). Dieses an der Universität des Saarlandes entwickelte Verfahren soll daher an dieser Stelle exemplarisch in seinen Ergebnissen kurz vorgestellt werden:
Ein Text (oder Dokument), z.B. ein Titel, ein Abstract, aber auch ein Brief oder eine Notiz werden maschinell auf CTX-Deskriptoren abgebildet. Dafür ein Beispiel: Bei der Eingabe einer Kurzbeschreibung aus einer Offenlegungsschrift des Deutschen Patentamts ermittelt das Verfahren folgende Deskriptoren:

Abb. 1a: Text (Auszug):

> *Eine Grabeinfassung hat einen viereckigen Rahmen, der aus zwei Laengssteinen (2) und zwei Quersteinen (2) besteht * Die Stirnflaechen der Quersteine (3) oder Laengssteine (2) liegen den Enden der Seitenflaechen der Laengssteine (2) oder Quersteine (2) gegenueber * Zwischen jede dieser Stirnflaechen (5) und die benachbarte Seitenflaeche (4) ist eine elastische Einlage (6) eingelegt **



3	1	3	7	0	0	6	BENACHBART	ADJ 00	
3	1	3	8	0	0	6	BENACHBARTE SEITENFLAECHE	SUB	A 7
1	1	1	15	2	2V6		BESTEHEN	VRB 00	
3	1	3	12	1	4N6		EINLAGE	SUB 20	
3	1	3	13	1	1V6		EINLEGEN	VRB 00	
3	1	3	11	1	4N6		ELASTISCH	ADJ 00	
3	1	3	12	1	4N6		ELASTISCHE EINLAGE	SUB	A 11
2	1	2	9	1	4N6		ENDE	SUB 30	
2	1	2	9	1	4N6		ENDE G SEITENFLAECHE	SUB	G 11
2	1	2	7	1	1V6		GEGENUEBERLIEGEN	VRB 00	
1	1	1	2	1	1N6		GRABEINFASSUNG	SUB 20 K	
2	1	2	6	1	1N6		LAENGSSTEIN K QUERSTEIN	SUB	KI 4
2	1	2	13	1	4N6		LAENGSSTEIN	SUB 10 K	
1	1	1	14	2	4N6		QUERSTEIN K LAENGSSTEIN	SUB	KI 11
1	1	1	14	2	4N6		QUERSTEIN	SUB 10 K	
1	1	1	6	1	2N6		RAHMEN	SUB 10	
1	1	1	6	1	2N6		RAHMEN HABEN	SUB	V 3
2	1	2	11	1	4N6		SEITENFLAECHE	SUB 20	
2	1	2	11	1	4N6		SEITENFLAECHE G QUERSTEIN	SUB	GE 15
2	1	2	11	1	4N6		SEITENFLAECHE G LAENGSSTEIN	SUB	G 13
2	1	2	2	1	1N6		STIRNFLAECHE	SUB 20 K	
2	1	2	2	1	1N6		STIRNFLAECHE G QUERSTEIN	SUB	G 4
2	1	2	2	1	1N6		STIRNFLAECHE G LAENGSSTEIN	SUB	GE 6
1	1	1	5	1	2N6		VIERECKIG	ADJ 00	
1	1	1	6	1	2N6		VIERECKIGER RAHMEN	SUB	A 5
1	1	1	11	2	4N6		ZWEI LAENGSSTEIN	SUB	A 10
1	1	1	14	2	4N6		ZWEI QUERSTEIN	SUB	A 13

Abb. 1 b: Deskriptoren

Legende: K = K-Relation (a in Konjunktion zu b)
G = G-Relation (b Genitiv-Attribut zu a)
Adjektiv-Nomen-Relationen stehen ohne Merkmal. Die Zahlen und Markierungen bezeichnen Quellen- und Textangaben (z.B. Satz- und Wortnummern).

Dabei werden Funktionswörter (hier z.B. DER, IST, JEDE, UND ...) eliminiert, die sinntragenden Wörter auf Grundformen reduziert, soweit sie flektiert sind:

Textwort	*Deskriptor*
LAENGSSTEINE	*LAENGSSTEIN*
SEITENFLAECHEN	*SEITENFLAECHE*
STIRNFLAECHEN	*STIRNFLAECHE*
QUERSTEINE	*QUERSTEIN*
VIERECKIGEN	*VIERECKIG*
EINGELEGT	*EINLEGEN*

Abb. 2: Wortformenreduktion (Auszug)

Die ermittelten mehrwortigen Begriffe (v.a. Adjektiv-Nomen-Verknüpfungen) bzw. Wortrelationierungen erlauben eine präzisere Deskribierung:

BENACHBARTE SEITENFLAECHE
ELASTISCHE EINLAGE

SEITENFLAECHE G LAENGSSTEIN
QUERSTEIN K LAENGSSTEIN

Abb. 3:Wortverknüpfungen (Auszug)

Aufgrund einer automatischen Permutierung der K-Relation (SYNK) und des automatischen Aufbaus einer Expansionsrelation (EXP) werden darüber hinaus dem Benutzer weitere Hilfen geboten. Mithilfe der Expansionsrelation kann er sich vor einer Präzisierung im Dialog über konkret vorhandene Präzisierungsmöglichkeiten informieren.

STEIN EXP QUERSTEIN
ELASTISCH EXP ELASTISCHE EINLAGE
VIERECKIG EXP VIERECKIGER RAHMEN
LAENGSSTEIN K QUERSTEIN SYNK QUERSTEIN K LAENGSSTEIN

Abb. 4 : Beispiele für (automatisch erstellte) textbezogene Synonym- und Expansionsrelationen (Auszug)

Die morphosyntaktischen Relationierungen konkretisieren einen Teil der bei einer entsprechenden automatischen Sprachanalyse ermittelten Strukturinformationen. Sie sollen gleichsam eine "gewohnte" Strategie beim Retrieval unterstützen.

Analog lassen sich (allerdings weitgehend intellektuell zu pflegende) morphologische Relationen (besonders wortklassenübergreifende Ableitungen, d. h. Derivationsrelationen) verwenden. Aus dem Textbeispiel sind (anwendungsspezifisch) u. a. folgende begriffliche Verknüpfungen zwischen Adjektiven (A), Verben (V) und Substantiven (S) ableitbar:

ELASTISCH	*DAS ELASTIZITÄT*
BENACHBART	*DAS NACHBAR*
BENACHBART	*DAS NACHBARSCHAFT*
EINLEGEN	*DVS EINLAGE*

Abb.5: Beispiele zu Derivationsbeziehungen

Ein zu CTX analoges Verfahren wurde in der DDR zum Indexieren medizinischer Befunde entwickelt /12/. Das Verfahren INDEX2 umfasst eine lexikalische Analyse mit Grundformermittlung, das Ausblenden von Trivialwörtern und den Aufbau von Deskriptorenketten, wobei darüber hinaus auch "semantische" kontextuelle Relationen, z.B. Negation und Konjunktivierung (i.S. von Modalangaben wie MÖGLICHERWEISE, WAHRSCHEINLICH) eingebracht werden. Auch hier werden nach Angaben des Systementwicklers - ähnlich zu CTX - ein- und mehrgliedrige Nominalphrasen ermittelt: "Die dominierende Rolle nimmt hierbei das Substantiv und seine attributive Umgebung in Form pränominaler Adjektiv- und Partizipialkonstruktionen sowie das Genitiv-Attribut ein" (S. 31). Vorläufer zu derartigen Verfahren finden sich bereits bei BRAUN und SEELBACH. Einfachere Mehrwort-Strategien (ohne explizite Verwendung von Lexika) sind auch bei der Gesellschaft fier Information und Dokumentation (GID) in Entwicklung /13/.

Andererseits muß man davon ausgehen, dass komplexe Verfahren wie die angegebenen Verknüpfungsmöglichkeiten an den Benutzer zunehmend höhere Anforderungen bzgl. der Systemhand-

habung stellen. Insofern bietet es sich an, automatische Retrievalfunktionen einzubringen. Wünschenswert sind, wie erwähnt, im Prinzip faktenorientierte Abfragen, wie sie in Frage-Antwort-Systemen Verwendung finden. Derartige Prozesse setzen jedoch hochentwickelte Analysetechniken und Wissensrepräsentationen voraus, die für den Einsatz in großen Informationssystemen heute noch nicht verfügbar sind /10/. Es bietet sich jedoch - gleichsam als Zwischenschritt - die Möglichkeit an, die sog. "Suchanfrage" wie ein Dokument zu bearbeiten. Im Rahmen von CTX wurde inzwischen ein derartiges Modul ("NATURA") entwickelt, das aus einer natürlichsprachigen Problembeschreibung analog zu den bei der Indexierung ermittelten ein- und mehrwortigen Deskriptoren eine entsprechende Deskriptorenliste aufbaut. Bei einer (vorgesehenen) Integration dieses Software-Bausteins in ein (bestehendes) Retrievalverfahren wird der Benutzer somit von der Aufgabe entlastet, die systembezogene normierte Form des Deskriptors selbst anzugeben. Er hat vielmehr die Möglichkeit, seine Problembeschreibung (z.B. in Form einer Phrase oder eines Satzes) anzugeben, woraufhin das System automatisch die entsprechenden Deskriptorketten aufbaut.

*Grabeinfassungen mit einem Rahmen, der elastische oder feste Einlagen besitzt **

GRABEINFASSUNG
EINFASSUNG
GRAB
RAHMEN
ELASTISCH
FEST
EINLAGE
BESITZEN
FESTE EINLAGE
ELASTISCHE EINLAGE
GRABEINFASSUNG P RAHMEN
EINLAGE BESITZEN

Abb. 5:Beispiel für "NATURA": Umsetzung einer natürlichsprachigen Suchanfrage in Deskriptoren

In einem weiteren Schritt könnte eine Reihe von Suchfragestrategien automatisch integriert werden, die es ermöglichen, direkt (unter Verwendung verschiedener sog. Ranking-Algorithmen) möglicherweise relevante Dokumente anzusprechen.

Während (flexions-)morphologische und syntaktische Analyseverfahren zumindest ausreichende Ergebnisse bereitstellen (Detailarbeit ist hierzu sicher noch erforderlich), ist der Bereich der Semantik, d. h. die Einbeziehung von Wort- oder Satzbedeutungen, bislang noch wenig erfolgreich automatisiert worden. Dies hängt auch damit zusammen, dass semantische Prozesse ohne Berücksichtigung von Morphologie und Syntax nur in Grenzen einsetzbar sind und zudem erheblich intensivere intellektuelle Investitionen (etwa bezüglich der Wörterbucherstellung) erfordern.

Mit Bezug auf die Erschließung und das Wiederfinden größerer Textmengen lassen sich jedoch zwei Entwicklungsmöglichkeiten erkennen:

- Semantische Disambiguierung, d. h. Vereindeutigung von Wörtern mit potentiell mehreren Bedeutungen;

- Einsatz im Bereich der Inhaltsanalyse (Content Analysis)

Wenn - zur Unterstützung eines Benutzers beim Retrieval - semantische Relationen von Begriffen ("semantisch" hier i.e.S.) einbezogen werden sollen, so wird ein unvereindeutigter Begriff ggf. zu kaum mehr vom Benutzer kontrollierbaren Textidentifikationen führen. Ein (zugegebenermaßen ungewöhnliches, wenn auch praktisch aufgetretenes) Beispiel dafür ist eine Recherche in einer JURIS-Datenbank mit den Begriffen BISCHOF und FIRMUNG. Sie führte zu einem Treffer, obgleich das Thema einen Unfall in Tauberbischofsheim und eine Firma (Plural: Firmen, Verb: firmen, Substantiv: Firmung) betraf. Man kann sich vorstellen, dass zunächst lange gerätselt wurde, warum gerade dieses einzige (zudem sehr umfangreiche) Dokument als Treffer ermittelt wurde. Eine Kombination aus "sinnloser" Zerlegung und Synonymrelation ohne "Bedeutungsdifferenzierung" führte zu diesem Ergebnis.
Vielfach wird als Argument gegen die Einführung einer Bedeutungsdifferenzierung vorgebracht, dass in einer Informationsbank in der Regel ja fachgebietsspezifische Daten stehen und ein Wort eben ggf. nur in einer Bedeutung auftrete. Wenn man unterstellen könnte, dass dies zutrifft, so bestünde eine automatische Differenzierungsregel eben nur in dem einfachen Hinweis, dass im Fachgebiet x ein Wort stets in der Bedeutung y zu vereindeutigen sei. (Es gibt jedoch genügend Belege, die diese Behauptung widerlegen.)

Eine weitere Behauptung, die in ihrer Wirkung weitaus problematischer ist, gründet sich darauf, dass durch eine Verknüpfung von Begriffen beim Retrieval solche Mehrdeutigkeiten von selbst eliminiert werden. Auch wenn man das BISCHOF-und-FIRMUNG-Beispiel zu den Ausnahmen rechnet, so stellt man doch damit das Prinzip der Begriffsverknüpfung auf den Kopf: Man präzisiert im Extremfall nur, um Mehrdeutigkeiten auszuscheiden, nicht aber, um das Thema sachlich einzugrenzen.

In jedem Falle ist ein semantische Disambiguierung erforderlich, wenn eine Indexierung mit dem zusätzlichen Ziel einer automatischen Sprachübersetzung (oder der Verknüpfung mit fremdsprachigen "Synonymen") verbunden werden soll. Dies ist v. a. für die Bearbeitung von Fachliteratur von Bedeutung.

Bei inhaltsanalytischen Verfahren wird der Bereich der Indexierung bzw. Verschlagwortung als Ausgangspunkt genommen. Bei dieser speziellen Art der Textanalyse wird, ausgehend von Textelementen (z.B. Wörtern), automatisch über eine Abbildung dieser Elemente auf ein Kategorien- oder Merkmalssystem und sich anschließenden (z.B. statistischen) Auswertungen versucht, Intentionen, Motive und ähnliches zu einem Text zu ermitteln. Im Zusammenhang mit Textarchivierung und -retrieval erscheinen damit einerseits thematische Aspektierungen möglich (z.B. "Übersichtsaufsatz", "Soziale Auswirkungen", "Evaluierungsproblematik"), andererseits lassen sich ganze Texte oder Textteile aufeinander beziehen (Clusteranalyse). Je differenzierter die sprachliche Grundlage für derartige (statistische) Prozesse ist, desto relevanter werden die Ergebnisse sein /11/.

4. Textverarbeitung im Büro

Was für den Bereich der Fachinformation halbwegs organisiert ist, erscheint in anderen Bereichen, in denen textuelle Informationen verarbeitet und archiviert werden, noch als eine unlösbare Aufgabe. Vor allem zu nennen ist hier der Bereich der betrieblichen Information (u. a. der Bürokommunikation), in dem tagtäglich Millionen von Briefen, Notizen, Berichten usf. erzeugt und abgelegt werden. Auch wenn der Wert dieser Information (z.B. im Hinblick auf die Kurzlebigkeit von Notizen) vielleicht mit anderen Maßstäben gemessen werden muß, so ist ihre Bedeutung für die innerbetriebliche Organisation wie für die außerbetriebliche Kommunikation wohl unumstritten. Als Beispiele für das besondere Interesse an einer Verbesserung durch elektronische Vermittlung können hier Electronic Mail (z.B. in ARPANET in den USA realisiert), TELETEXT und die Individualkommunikation bei Bildschirmtext (BTX) genannt werden. Eine intellektuelle Verdichtung bzw. Aufbereitung i.S. der Dokumentation (Abstracting, Indexing) erscheint von vornherein ausgeschlossen.

Mit der angehenden Entwicklung neuer Kommunikationstechnologien, insbesondere der Telematik (Breitbandkommunikationsnetze) und der Microprozessortechnik, aber auch der Bereitstellung von Speichertechniken größten Ausmaßes wird sich somit der Wandel vollziehen, der auf die Strategien der Archivierung und des Wiederfindens von textuellen Daten von entscheidendem Einfluss sein wird. Dabei müssen v.a. neuere Verfahren herangezogen werden, die auf linguistischen, d.h. sprachwissenschaftlichen Erkenntnissen beruhen und die sich zugleich auf große Textmengen relativ problemlos anwenden lassen.

Die Anwendung automatischer Verfahren der Erschließung und des Wiederfindens von textueller Information im Büro wird unter Berücksichtigung der aufgeführten Verfahren und Probleme von folgenden Kriterien abhängen:

(1) Robustheit und Integrationsfähigkeit der Texterschließungs- und -retrievalsoftware

Schriftlich fixierte "Dokumentmengen" im Büro - in seiner vagen, allgemeinen Bedeutung - sind z.B. die Notiz, der Brief, das Protokoll, der Bericht, die Adresse, der Terminkalender. Betriebsintern lassen sich manche derartige Texttypen (sofern bereits durch Textsysteme usf. entsprechende Unterstützung bzw. Bedienerführung gegeben ist) über Formularerfassung oder Menütechnik stärker formalisieren. Ein Großteil der Ablage- und Abfragemöglichkeiten wird sich daher - analog zu den Anfragen in Literaturinformationssystemen - an diesen strukturierbaren Teilen (z.B. ADRESSAT, ABSENDEDATUM, AKTENZEICHEN, BEZUG) orientieren. Soweit Bürokommunikationssysteme bereits freitextbezogene Funktionen anbieten werden, steht - ebenfalls in Analogie zur Fachinformation - zu erwarten, dass sie - aus Kosten- oder auch Pflegeerwägungen heraus - zeichenorientiert arbeiten werden, d, h. mit Methoden der Trunkierung und des Abstandsmaßes. Soweit wörterbuchbezogene Funktionen eingebracht werden (etwa im Sinne der Ermittlung von Grundformen, von Teilwörtern bei Komposita oder auch von Ableitungen), ist die Pflegekomponente zu bedenken. Allerdings stellen sich im Bürobereich derartige Probleme bereits auf (scheinbar) trivialerer Ebene, nämlich bei der automatischen Silbentrennung und der Rechtschreibkorrektur. Bei flexionsreichen Sprachen wie dem Deutschen, aber auch angesichts der Unregelmäßigkeiten bzw. Komplexität der Silbentrennung (wie im Englischen) bietet sich von vornherein der Aufbau und die Einbeziehung von Wörterbüchern an. Dies erfolgt - meist aufgrund vordergründiger Kostenüberlegungen, aber auch in Unkenntnis der Probleme - vielfach

in der Büroautomation noch eher ad-hoc und unsystematisch (man glaubt häufig, die Probleme rein "quantitativ" anhand von Wortformenlisten bewältigen zu können).

Das langsam einsetzende Umdenken in diesem Bereich wird jedoch bald zu systematischeren Verfahren führen.. So koppelt IBM derzeit in seinen Schreibsystemen ein Stammwörterbuch mit noch relativ oberflächigen Flexionsangaben mit der Silbentrennung. RANK-XEROX hat in Palo Alto ein Verfahren zur technischen Anbindung von Lexikonsystemen für Mikroprozessoren in Entwicklung, NIXDORF wird ebenfalls im Bereich Bürokommunikation in naher Zukunft mit systematisch-lexikonorientierten Verfahren für Rechtschreibhilfen und Silbentrennung auf dem Markt sein. Von dieser Stufe aus ist der Weg zur Implementierung von Freitextsuchverfahren (unter Integration von Silbentrennung, Rechtschreibhilfen und "phonetisiertem" Zugriff) nicht mehr allzu weit. Derartige Komponenten lassen sich ziemlich "robust" entwickeln, da jedem Kunden üblicherweise die Möglichkeit gegeben wird, fehlende Einträge zumindest zwischenzeitlich in "Kundenwörterbüchern" zu speichern. Sobald einmal die Techniken einer "Ferndiagnose" allgemein genutzt werden können, wird auch eine anwenderspezifische Systempflege durch den Systementwickler erleichtert. Umgekehrt müssen lexikalische Systeme - auch wenn sie auf Archivierungsprobleme ausgerichtet sind - in der Lage sein, Fragen der Silbentrennung und Rechtschreibung in integrierter Form mit zu berücksichtigen.

(2) Anpassung "höherwertiger" Verfahren von Großrechner- auf Mikroprozessor-Anwendungen

Auf Großrechnerebene existieren heute eine Reihe von Verfahren von der automatischen Sprachanalyse bis zur maschinellen Übersetzung, die bereits über ein reines Laborstudium hinausgekommen sind und im Sinne der vorliegenden Zielsetzung "intelligente" Verfahren einschließen. Zu nennen sind in diesem Zusammenhang wiederum das CTX-System (für die automatische Indexierung), aber auch Übersetzungssysteme wie SYSTRAN. Es ist bezeichnend, dass inzwischen für die Anwendung in Übersetzungsbüros (d.h. zur Unterstützung der intellektuellen Übersetzung) spezielle Systeme in Erprobung sind (hier sei auf das WEIDNER-System verwiesen; bei SYSTRAN wird in dem Anwendungstest bei der Kommission der Europäischen Gemeinschaften ein Textsystem - hier WANG - mit dem Großrechner gekoppelt). Aufgrund der wachsenden Leistungsfähigkeit der "Micros" wird die Umsetzung bzw. angepasste Neuentwicklung von Übersetzungsfunktionen weiter vorangetrieben werden. Auf diese Weise werden Analyseergebnisse zur Satzstrukturierung und Bedeutungsdifferenzierung verfügbar, die zu Archivierungs- und Retrievalzwecken herangezogen werden können.

Ein Problem, das sich bereits heute in diesem Zusammenhang stellt, ist - besonders bei großen Datenmengen - die Unterscheidung von im Text- oder Themenzusammenhang relevanten (Text-)Begriffen gegenüber weniger relevant erscheinenden Begriffen. Setzt man einmal voraus, dass auch im Microprozessor-Bereich Rechenzeitverhalten und Speicherplatzbedarf nicht mehr die entscheidende Rolle spielen (dies verschiebt die Realisierung solch komplexer Anwendungen noch in die - wenn auch absehbare - Zukunft), so werden v.a. mathematische Clusterverfahren und Prozeduren analog zur Inhaltsanalyse hierfür an Bedeutung gewinnen.

(3) Benutzerfreundliche Retrievalschnittstelle

Es ist einer Sekretärin, aber auch einem Sachbearbeiter oder dem Manager (gleichsam als den typischen Vertretern bzw. Partnern in einer Bürokommunikation) nicht zuzumuten (zumindest im Sinne einer geringeren Akzeptanz), dass sie sich irgendwelcher technischer Tricks oder "tieferer"

sprachwissenschaftlicher Kenntnisse bedienen, um textbezogene Informationen mit ausreichender Präzision zu recherchieren. (Es ist auch eine Frage, ob man dies dem "studierten" Wissenschaftler in der Fachinformation zumuten sollte.)

Die einfachste Schnittstelle ist - neben dem faktenorientierten Frage-Antwort-System, das m.E. (wie erwähnt) vorerst für eine allgemeine Anwendung noch wenig praktikabel erscheint - die referenzorientierte natürlichsprachige Problembeschreibung. Sie ist eingebettet in einen Recherchedialog Mensch-Computer, wobei das System aufgrund von themenbezogenen Termini, die in der natürlichsprachigen Problembeschreibung auftreten, aufgrund von (thesaurusartigen) Begriffsvernetzungen und aufgrund von mathematischen "Relevanzfunktionen", die aus der Struktur der Problembeschreibung im Black-Box-Prinzip (d.h. in der Strategie für den Benutzer unzugänglich) abgeleitet werden und die "Nähe" der Problembeschreibung zu gespeicherten Akten, Notizen, Berichten (bzw. Ausschnitten dazu) ermittelt, evtl. geordnet nach dem Grad dieser "mathematisch-linguistisch" errechneten Nähe /16/.

Verfahren dieser Art setzen nach aller bisherigen Erkenntnis weit mehr voraus als eine Reduktion von Wörtern eines Textes auf Grundformen. Umgekehrt ist bereits eine Vielzahl relevanter Informationen zu Begriffsbeziehungen und -vernetzungen in vorhandenen "gedruckten", z.T. auch maschinenlesbar gespeicherten Lexika, Enzyklopädien, Terminologien und Thesauri fixiert, die die Entwicklung solcher Systeme praktikabel und wirtschaftlich erscheinen lässt. Der Markt, den solche Indexierungs- und Archivierungssysteme vorfinden werden, wird auch das nicht unbeträchtliche Investitionsvolumen rechtfertigen, das für den Aufbau und die Pflege höherwertiger Verfahren der sprachbezogenen Texterschließung- und -verarbeitung im Büro erforderlich zu sein scheint. Vielleicht kann durch eine zukunftsorientierte Unternehmensentscheidung bei den Entwicklern und Anbietern derartiger Bürokommunikationssysteme dem Benutzer der Umweg erspart bleiben, der im Bereich der Fachinformation zu unhandlichen, oberflächenorientierten Verfahren geführt hat, die zudem wegen der inzwischen damit aufbereiteten Datenmengen zu entsprechenden Ablöseschwierigkeiten führen. Jedenfalls sollte es gelingen, auf dem Weg zu intelligenteren Verfahren im Bereich der Bürokommunikation Sackgassen zu vermeiden, die sich bei Ad-hoc-Lösungen in der Sprachdatenverarbeitung ergeben.

ANMERKUNGEN UND LITERATUR

/1/ WINGERT, F., (Hrsg.) Klartextverarbeitung. Medizinische Informatik und Statistik. Berlin, Heidelberg, New York 1978. ISBN 3-540-8634-X.

/2/ WINGERT, F.: Klartextverarbeitung in der Medizin, In: Klartextverarbeitung (vgl. /1/), S. 1.

/3/ Ebda. S. 1. /4/ Ebda. S.1f.

/5/ "Freitext" bedeutet in dieser Unterscheidung die völlig freie Formulierung eines Textes in einer natürlichen Sprache; beim "Klartext" können - mit Bezug auf ein vorgegebenes Auswertungsverfahren - die Regeln der natürlichen Sprache (leicht) eingeschränkt bzw. um (fachgebiets-)spezifische Elemente erweitert sein. Die Unterscheidung ist im vorliegenden Zusammenhang von untergeordneter Bedeutung, da in komplexen Systemen mit variierenden Benutzern - auf die letztlich ein Textarchivierungssystem ausgerichtet sein muß - eine Einhaltung reduzierter Sprach-

regeln wenig sinnvoll erscheint und ohne entsprechenden Schulungsaufwand nicht sichergestellt werden kann.

/6/ In der Linguistik hat man für den Teil des sprachlichen Codes, der allen Sprachteilnehmern "idealerweise" gemeinsam ist, den Begriff des "denotativen Codes" geprägt. Entsprechend ist der Teil, der von der Erfahrung des Einzelnen abhängt, mit "konnotativer Code" bezeichnet. Analoges gilt sicherlich auch für die Fachsprachen.

/7/ Vgl.: PASSAT: Automatische Selektion von Stichwörtern aus Texten. PASSAT BS2000 Verfahrensbeschreibung, München (Siemens) 1980.

/8/ Zu CTX vgl. weiter unten.

/9/ STAIRS: Storage And Information Retrieval System (Dokumentations- und Retrieval-System der IBM). TLS: Thesaurus and Linguistic Integrated System (Thesaurus- und Flexionsformen-System der IBM).

/10/ Erfolgversprechende Ansätze sind inzwischen auch in der Bundesrepublik Deutschland im Laborstadium. Hierzu sei auf die Verfahren BEAST/BACON (Berlin), HAM/RPM (Hamburg), USL (IBM Heidelberg) oder OBJTALK (Stuttgart) verwiesen. Ein Test an großen Datenmengen steht allerdings hierzu noch aus.

/11/ Bei CTX werden gegenwärtig beide Ansätze verfolgt. In Erprobung ist eine automatische Indexierung unter Bedeutungsdifferenzierung mit Abstracts aus Veröffentlichungen des Wissenschaftszentrums Berlin (WZB); mit Bezug auf die Inhaltsanalyse werden gegenwärtig verschiedene deutsche inhaltsanalytische Wörterbücher in einer Vorstufe mit dem CTX-Wörterbuch verknüpft.

Literatur zu CTX: CTX - Ein Verfahren zur computergestützten Texterschließung. Vorgesehen zur Veröffentlichung in der BMFT-Reihe "Information und Dokumentation" (BMFT-ID).

/12/ Vgl.: GRAICHEN, D:: Thesaurusunabhängiges Indexieren medizinischer Befunde mit "INDEX 2". In: Informatik 28 (1981) 4 , S. 30-35.

/13/ Vgl. BRAUN, S.: Automatische Indexierung durch linguistische Syntaxanalyse. GI-Jahrestagung (1973). Berlin, S. 414-420.

/14/ SEELBACH, D.: Computerlinguistik und Dokumentation. Key Phrases in Dokumentationsprozessen. München 1975.

/15/ ROSTEK, L.: Methoden des partiellen Parsing für das automatische Indexing - Syntaxgraphen zur Analyse von Sprachmustern. In: R. Kuhlen (ed.) Datenbasen, Datenbanken, Netzwerke, Bd. 1, München 1979, S. 251-282.

/16/ Vgl. hierzu auch die Untersuchungen zum CONDOR-System, v.a.: Wieland, U.: Recherche auf. der Basis einer syntaxorientierten maschinellen Sprachanalyse. Diss. Regensburg 1979.

Anlagen

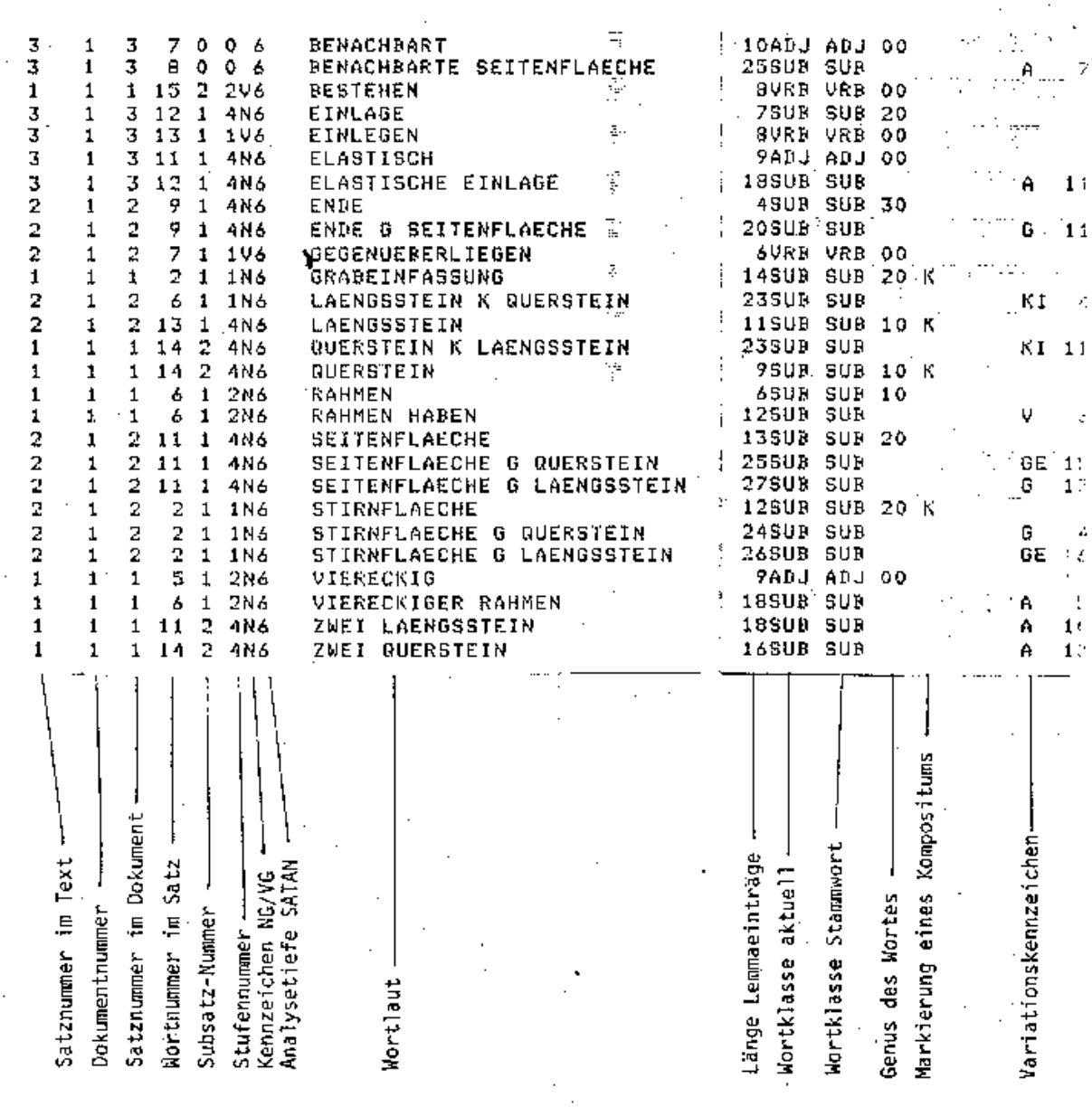
COMPUTERGESTUTZTE TEXTERSCHLIESSUNG mit CTX

dargestellt am Beispiel eines Textes des Deutschen Patentamts München

OFFENLEGUNGSSCHRIFT Nr. DE 31 00 360 Al

> *Eine Grabeinfassung hat einen viereckigen Rahmen , der aus zwei Laengssteinen (2) und zwei Quersteinen (3) besteht * Die Stirnflaechen der Quersteine (3) oder Laengssteine (2) liegen den Enden der Seitenflaechen der Laengssteine (2) oder Quersteine (2) gegenueber * Zwischen jede dieser Stirnflaechen (5) und die benachbarte Seitenflaeche (4) ist eine elastische Einlage (6) eingelegt **

DESKRIBIERUNG



DERIVATIONSRELATIONEN

ANATOMIE	DSA	ANATOMISCH
AUSBILDUNG	DSV	AUSBILDEN
BEDEUTUNG	DSV	BEDEUTEN
DURCHFUEHRUNG	DSV	DURCHFUEHREN
	DSA	DURCHFUEHRBAR
ENTWICKLUNG	DSV	ENTWICKELN
ERLAEUTERUNG	DSV	ERLAEUTERN
GESTALTUNG	DSV	GESTALTEN
MODELL	DSA	MODELLHAFT
MOEGLICHKEIT	DSA	MOEGLICH
SIMULATION	DSV	SIMULIEREN

SYNONYMIE ATTRIBUT - KOMPOSITUM
(intellektuelle Pflege, z.T. auf Systemvorschlag)

BEDEUTUNG	G SIMULATION	SYN	SIMULATIONSBEDEUTUNG
MOEGLICHKEIT	G SIMULATION	SYN	SIMULATIONSMOEGLICHKEIT

ZERLEGUNG DER KOMPOSITA

Kompositumstamm — Zerlegung

GRABEINFASSUNG — /GRAB/EINFASSUNG
1. Wortklasse = SUB Wortlaut = EINFASSUNG
2. Wortklasse = SUB Wortlaut = GRAB

LAENGSSTEIN — +LAENGS/STEIN
1. Wortklasse = SUB Wortlaut = STEIN
2. Wortklasse = PRF Wortlaut = LAENGS

QUERSTEIN — /QUER/STEIN
1. Wortklasse = SUB Wortlaut = STEIN
2. Wortklasse- VRB Wortlaut = QUER

STIRNFLAECHE — /STIRN/FLAECHE
1. Wortklasse = SUB Wortlaut = FLAECHE
2. Wortklasse = SUB Wortlaut = STIRN

DERIVATIONSRELATIONEN

BENACHBART	DAS	NACHBAR
EINLAGE	DSV	EINLEGEN
ELASTISCH	DAS	ELASTIZITAET
ENDE	DSV	ENDEN
VIERECKIG	DAS	VIERECK

SYNONYMIE ATTRIBUT - KOMPOSITUM
(intellektuelle Pflege, z.T. auf Systemvorschlag)

ENDE	G SEITENFLAECHE	SYN	SEITENFLAECHENENDE

COMPUTERGESTUTZTE TEXTERSCHLIESSUNG mit CTX
dargestellt am Beispiel eines Textes des Wissenschaftszentrum Berlin (WZB)
TEXT NR. A-00393

$TXTA-00393
$DOK
Die vorliegende Arbeit beschaeftigt sich mit den regional unterschiedlich hohen Behindertenanteilen in der Wohnbevoelkerung und versucht diese ueber die Belastungsfelder des Erwerbslebens und der Umwelt zu erklaeren Eine multiple Regressionsanalyse*

*kommt zu dem Ergebnis, dass der Anteil der Beschaeftigten in besonders belasteten Wirtschaftszweigen und die Bevoelkerungsdichte 39 Prozent der Varianz der regionalen Behindertenpopulation erklaeren koennen**
Zur Durchfuehrung der Analyse war eine Aggregation der Behindertenzahlen, die nach Kreisen vorlagen, auf die Ebene der Arbeitsamtsbezirke notwendig Es werden hier erstmals Behindertenzahlen fuer die Arbeitsamtsbezirke vorgelegt**

DESKRIBIERUNG:

Satznummer in Text	Dokumentnummer Satznummer im Dok.	Wortnummer im Satz	Subsatz-Nummer Stufennummer Kennzeichen Ist/VB Analysetiefe Stamm Bedeutungsnummer	Wortlaut	Länge Lemmaeintr. Wortklasse aktuell	Wortklasse Stammw.	Genus d.Wortes attributives Part. Variationskennz.
3	1 3	7	1 3N7	AGGREGATION	11SUB	SUB	20
3	1 3	7	1 3N7	AGGREGATION G BEHINDERTENZAHL	29SUB	SUB	G
3	1 3	4	1 1N7	ANALYSE	7SUB	SUB	20
2	1 2	11	2 4N7	ANTEIL	6SUB	SUB	10F
2	1 2	11	2 4N7	ANTEIL G BESCHAEFTIGTE	22SUB	SUB	G
2	1 2	11	2 4N7	ANTEIL K BEVOELKERUNGSDICHTE	28SUB	SUB	K
1	1 1	3	1 1N7	ARBEIT	6SUB	SUB	20F
3	1 3	20	1 5N7	ARBEITSAMTSBEZIRK	17SUB	SUB	10
2	1 2	27	2 8N7	BEHINDERTENPOPULATION	21SUB	SUB	20
4	1 4	5	1 4N7	BEHINDERTENZAHL VORLEGEN	24SUB	SUB	V
1	1 1	11	1 5N7	BEHINDERTENANTEIL	17SUB	SUB	10
3	1 3	9	1 3N7	BEHINDERTENZAHL	15SUB	SUB	20
2	1 2	16	2 4N7	BELASTEN	8ADJ	VRB	02
2	1 2	17	2 4N7	BELASTETER WIRTSCHAFTSZWEIG	27SUB	SUB	A
1	1 1	20	2 8N7	BELASTUNGSFELD	14SUB	SUB	30
1	1 1	20	2 8N7	BELASTUNGSFELD G ERWERBSLEBEN	29SUB	SUB	G
1	1 1	20	2 8N7	BELASTUNGSFELD G UMWELT	23SUB	SUB	GE

Satznummer im Text	Dokumentennummer	Satznummer im Dok.	Wortnummer im Satz	Substantivnummer	Stufennummer	Kennzeichen Haupt- Analysegrundlage Satzan	Bedeutungsnummer	Wortlaut	Länge Lemmaeintr.	Wortklasse aktuell	Wortklasse Stamm	Genus d.Kurzes attributives Part.	Variationskennz.
2	1	2	13	2	4N7			BESCHAEFTIGTE P WIRTSCHAFTSZWEIG	32SUB	SUB			P
1	1	1	4	1	1V7			BESCHAEFTIGEN	13VRB	VRB	00		
2	1	2	13	2	4N7			BESCHAEFTIGTE	13SUB	SUB	20		
2	1	2	20	2	4N7			BEVOELKERUNGSDICHTE K ANTEIL	28SUB	SUB			KI
2	1	2	20	2	4N7			BEVOELKERUNGSDICHTE	19SUB	SUB	20		
3	1	3	2	1	1N7			DURCHFUEHRUNG	13SUB	SUB	20F		
3	1	3	2	1	1N7			DURCHFUEHRUNG G ANALYSE	23SUB	SUB			G
3	1	3	18	1	5N7			EBENE	5SUB	SUB	20F		
3	1	3	18	1	5N7			EBENE G ARBEITSAMTSBEZIRK	25SUB	SUB			G
2	1	2	7	1	2N7			ERGEBNIS	8SUB	SUB	30		
1	1	1	27	2	2V7			ERKLAEREN	9VRB	VRB	00		
2	1	2	28	2	2V7			ERKLAEREN KOENNEN	17VRB	VRB			M
1	1	1	27	2	2V7			ERKLAEREN VERSUCHEN	19VRB	VRB			M
1	1	1	22	2	8N7			ERWERBSLEBEN	12SUB	SUB	30		
1	1	1	22	2	8N7			ERWERBSLEBEN K UMWELT	21SUB	SUB			K
1	1	1	10	1	5N7			HOCH	4ADJ	ADJ	00		
1	1	1	11	1	5N7			HOHER BEHINDERTENANTEIL	24SUB	SUB			A
3	1	3	13	2	9N7			KREIS	5SUB	SUB	10		
2	1	2	2	1	1N7			MULTIPEL	8ADJ	ADJ	00		
2	1	2	3	1	1N7			MULTIPLE REGRESSIONSANALYSE	28SUB	SUB			A
3	1	3	21	0	0 7			NOTWENDIG	9ADJ	ADJ	00		
2	1	2	22	2	8N7			PROZENT	7SUB	SUB	30		
2	1	2	22	2	8N7			PROZENT G VARIANZ	17SUB	SUB			G
1	1	1	8	0	0 7			REGIONAL	8ADJ	ADJ	00		
2	1	2	27	2	8N7			REGIONALE BEHINDERTENPOPULATION	31SUB	SUB			A
2	1	2	3	1	1N7			REGRESSIONSANALYSE	18SUB	SUB	20		
1	1	1	25	2	8N7			UMWELT	6SUB	SUB	20		
1	1	1	25	2	8N7			UMWELT K ERWERBSLEBEN	21SUB	SUB			KI
1	1	1	9	0	0 7			UNTERSCHIEDLICH	15ADJ	ADJ	00		
2	1	2	24	2	8N7			VARIANZ	7SUB	SUB	20		
2	1	2	24	2	8N7			VARIANZ G BEHINDERTENPOPULATION	31SUB	SUB			G
1	1	1	16	2	2V7			VERSUCHEN	9VRB	VRB	00		
4	1	4	9	1	1V7			VORLEGEN	8VRB	VRB	00		
1	1	1	2	1	1N7			VORLIEGEN	9ADJ	VRB	01		
1	1	1	3	1	1N7			VORLIEGENDE ARBEIT	18SUB	SUB			A
2	1	2	17	2	4N7			WIRTSCHAFTSZWEIG	16SUB	SUB	10		
1	1	1	14	1	6N7			WOHNBEVOELKERUNG	16SUB	SUB	20		
2	1	2	21	2	8N7			39	2NUM		00		
2	1	2	22	2	8N7			39 PROZENT	12SUB	SUB			A

ZERLEGUNG DER KOMPOSITA:

Kompositumstamm	Zerlegung
ARBEITSAMTSBEZIRK	/ARBEITSAMT*S/BEZIRK

1. Wortklasse = SUB Wortlaut = BEZIRK
2. Wortklasse SUB Wortlaut = ARBEITSAMT

BEHINDERTENANTEIL /BEHINDERTE*N/ANTEIL.

1. Wortklasse = SUB Wortlaut = ANTEIL
2. Wortklasse = SUI Wortlaut .= BEHINDERTE

BEHINDERTENPOPULATION /BEHINDERTE*N/POPULATION
1. Wortklasse = SUB Wortlaut = POPULATION
2. Wortklasse = SUB Wortlaut = BEHINDERTE
BEHINDERTENZAHL /BEHINDERTE*N/ZAHL
1. Wortklasse = SUB Wortlaut = ZAHL
2. Wortklasse = SUB Wortlaut = BEHINDERTE
BELASTUNGSFELD /BELASTUNG*S/FELD
1. Wortklasse = SUB Wortlaut = FELD
2. Wortklasse = SUB Wortlaut = BELASTUNG
REGRESSIONSANALYSE /REGRESSION*S/ANALYSE
1. Wortklasse = SUB Wortlaut = ANALYSE
2. Wortklasse= SUB Wortlaut = REGRESSION
WIRTSCHAFTSZWEIG /WIRTSCHAFT*S/ZWEIG
1. Wortklasse = SUB Wortlaut = ZWEIG
2. Wortklasse = SUB Wortlaut = WIRTSCHAFT

DERIVATIONSRELATIONEN:

DSV	AGGREGATION	AGGREGIEREN
DSV	ANALYSE	ANALYSIEREN
DSV	ARBEIT	ARBEITEN
DSV	BELASTUNG	BELASTEN
DSA	BELASTUNG	BELASTBAR
DVS	BESCHAEFTIGEN	BESCHAEFTIGUNG
DSV	DURCHFUEHRUNG	DURCHFUEHREN
DVS	ERKLAEREN	ERKLAERUNG
DAS	NOTWENDIG	NOTWENDIGKEIT
DSV	VARIANZ	VARIIEREN
DVS	VORLIEGEN	VORLAGE

SYNONYMIE ATTRIBUT_- KOMPOSITUM:
(intellektuelle Pflege, z.T. auf Systemvorschlag)

ANTEIL G BESCHAEFTIGUNG	SYN	BESCHAEFTIGUNGSANTEIL
BEHINDERTENANTEIL	SYN	ANTEIL G BEHINDERTE
BEHINDERTENZAHL	SYN	ZAHL G BEHINDERTE
ARBEITSAMTSBEZIRK	SYN	BEZIRK G ARBEITSAMT

COMPUTERGESTÜTZTE TEXTERSCHLIESSUNG mit CTX

dargestellt am Beispiel eines Textes
der Bundesanstalt für Materialprüfung Berlin
(Text Nr. D18200146)

Bedeutung der Simulation bei Gestaltung und Erlaeuterung von Fertigungssystemen *

*Erlaeuterung der Anatomie eines Fertigungssystems * Moeglichkeiten der Simulation von Fertigungssystemen * Durchfuehrung der Software-Simulation * Simulation mit. funktionalen Modellen Entwicklung des Pilot-Hardware-Systems * Ausbildung an Fertigungssystemen **

CTX-DESKRIBIERUNG

Satznummer im Text	Dokumentnummer	Satznummer im Dokument	Wortnummer im Satz	Subsatz-Nummer	Stufennummer	Kennzeichen NG/VG	Analysetiefe SATAN	Wortlaut	Wortklasse Stammwort	Genus des Wortes	Variationskennzeichen
2	1	2	3	1	1N7			ANATOMIE	SUB	20	0 0 0
2	1	2	1	1	1N7			ANATOMIE FERTIGUNGSSYSTEM G	SUB		
7	1	7	1	1	1N7			AUSBILDUNG	SUB	20	2 0 0
1	1	1	1	1	1N7			BEDEUTUNG	SUB	20	2 0 0
1	1	1	9	1	3N7			BEDEUTUNG SIMULATION G	SUB		
4	1	4	1	1	1N7			DURCHFUEHRUNG	SUB	20	2 3 0
4	1	4	3	1	1N7			DURCHFUEHRUNG SOFTWARE-SIMULATION G	SUB		
6	1	6	1	1	1N7			ENTWICKLUNG	SUB	20	2 0 0
6	1	6	3	1	1N7			ENTWICKLUNG PILOT-HARDWARE-SYSTEM G	SUB		
2	1	2	1	1	1N7			ERLAEUTERUNG	SUB	20	0 0 0
2	1	2	5	1	1N7			ERLAEUTERUNG ANATOMIE G	SUB		
1	1	1	1	1	1N7			ERLAEUTERUNG FERTIGUNGSSYSTEM G	SUB		
3	1	3	5	1	1N7			FERTIGUNGSSYSTEM	SUB	30	0 0 0
5	1	5	3	1	2N7			FUNKTIONAL	ADJ	00	0 0 0
5	1	5	4	1	2N7			FUNKTIONALES MODELL	SUB		
1	1	1	5	1	3N7			GESTALTUNG FERTIGUNGSSYSTEM G	SUB		
1	1	1	5	1	3N7			GESTALTUNG	SUB	20	0 0 0
1	1	1	7	1	3N7			GESTALTUNG ERLAEUTERUNG K	SUB		
5	1	5	4	1	2N7			MODELL	SUB	30	0 0 0
3	1	3	1	1	1N7			MOEGLICHKEIT	SUB	20	0 0 0
3	1	3	5	1	1N7			MOEGLICHKEIT SIMULATION G	SUB		
6	1	6	3	1	1N7			PILOT-HARDWARE-SYSTEM	SUB	30	0 0 0
1	1	1	3	1	1N7			SIMULATION	SUB	20	0 0 0
3	1	3	1	1	1N7			SIMULATION FERTIGUNGSSYSTEM G	SUB		
4	1	4	3	1	1N7			SOFTWARE-SIMULATION	SUB	20	0 0 0