Inhalt

Seite

Regionalkonzepte

Vorwort

Die Sektion Klartextanalyse entstand 1972 als Untergruppe der Arbeitsgruppe für Medizinische Informationsverarbeitung in der Deutschen Gesellschaft für Medizinische Dokumentation und Statistik. Unser Arbeitsziel war, die bisher im deutschen Sprachbereich erarbeiteten Textauswertungsverfahren zu analysieren und eine Zwischenbilanz des theoretischen und praktischen Entwicklungsstandes auf diesem Gebiet zu erstellen. Nachdem zunächst die unmittelbaren Arbeitsprobleme der einzelnen Sektionsmitglieder in mehreren Arbeitssitzungen abgeklärt worden waren, war es jetzt möglich und erforderlich, den Entwicklungsstand in einem Symposium zusammenzufassen.

Wir haben Wien als Tagungsort gewählt, weil sich dort innerhalb des letzten Jahrzehntes ein Schwerpunkt der medizinischen Informatik im deutschen Sprachgebrauch entwickelt hat und wir uns deshalb besondere Resonanz versprachen. Die Veranstaltung wurde ermöglicht durch die Einladung von Herrn Prof.Dr.J.H. Holzner, der Mitglied unserer Sektion ist und der die organisatorische Durchführung in seinem Institut übernommen hat.

Der Tagungsinhalt gliederte sich in zwei Schwerpunkte, und zwar in einen allgemeinen und einen speziellen Teil. Im allgemeinen Teil wurden Probleme der Semantik, der Informationstheorie sowie der Definition der Klartextanalyse behandelt, im speziellen Abschnitt stand neben einem Überblick über ausländische Systeme vorerst die Vorstellung und die erste Anwendung des Thesaurus der Arbeitsgemeinschaft Klartextanalyse (AGK-Thesaurus) im Vordergrund.

Eine Bilanz deutschsprachiger Systeme ohne das Verfahren des Grazer Arbeitskreises (Becker, Gell, Muhri) wäre unvollständig. Durch Umstände, die sich dem Einfluß der Veranstalter entzogen, konnte jenes Verfahren nicht abgehandelt werden. Die Mitarbeit einiger Sektionsmitglieder in der Arbeitsgruppe PRATT hat zusätzliche Schwerpunkte für die zukünftige Sektionsarbeit ergeben. Es ist zu erwarten, daß die Darstellung des Grazer Systems sich in einer der nächsten Sitzungen, die sich vorwiegend mit den Möglichkeiten und Problemen des Pratt'schen Systems befaßt, noch besser integriert werden kann.

Die Sektion Klartextanalyse wird in der Erfüllung ihrer Aufgabe - dem Erzielen weiterer Fortschritte in der Auswertung medizinischer Befundtexte - sehr davon abhängen, wie weit die bisher erarbeiteten Methoden und praktischen Erkenntnisse weiterhin genutzt werden und verfügbar bleiben. Durch das Engagement der Fa. Siemens ist es ermöglicht worden, die bisherigen Referate in diesem Symposiumsbericht in vollem Wortlaut mitzuteilen. Damit ist die künftige Arbeit unserer Sektion wesentlich gefördert worden.

Wichtig für die Arbeit der Sektion war auch die zahlreiche Teilnahme von Mitgliedern und Nichtmitgliedern. Eine intensive und lebendige Diskussion, wie sie in Wien stattfand, ist für die weitere Arbeit unserer Sektion unerläßlich. In diesem Sinne ist die Veröffentlichung der Referate als Fortsetzung eines Dialoges zu verstehen.

Röttger Ertel Feigl 3.11.1973

Semantische Aspekte medizinischer Informationen

M. Ertel

Forschungslaboratorium der Siemens AG

In der gegenwärtigen und zukünftigen medizinischen Versorgung hat die Behandlung und Weitervermittlung von Informationen eine eminente Bedeutung. Es besteht in diesem Zusammenhang nicht nur die Notwendigkeit, vorhandenes Wissen in Formen zu packen, die besser handhabbar sind als bisher. Ich sehe auch eine Reihe von Problemstellungen, wo es nötig sein wird, bildliche Informationen und den Formenreichtum der Sprache dem Kommunikationsprozess in der Medizin in neuer Form zu erschließen. Über einige Untersuchungen zu diesen Fragenkomplexen wird referiert.

Die Identifizierung von Bildern eines Bildarchivs mit Computerunterstützung

Die Informationsübermittlung per Bild hat im Kommunikationsprozess der Menschen nicht die große Bedeutung wie das Wort. Trotzdem gibt es zahlreiche Fälle, wo das Bild wegen seines großen Informationsgehaltes nicht entbehrt werden kann. Eine bedeutende Aufgabe ist es, bildliche Informationen so zu archivieren, daß man auf sie im Bedarfsfalle wieder schnell zurückgreifen kann. Das System, das ich für meine Ausführungen im Auge habe, befaßt sich mit der Identifizierung von Gesichtern, deren Bilder in einem Archiv aufbewahrt sind [1]. Es setzt einen Computerdialog ein, um ein bestimmtes Bild oder ähnliche Bilder aus dem gesamten Bild-Korpus herauszufinden. Die Untersuchungen wurden in einem technischen Labor gemacht, die Ergebnisse sind auf ähnlich gelagerte Fälle in der Medizin unmittelbar übertragbar.

Folgende Teilkomplexe sind in diesem Zusammenhang abzuhandeln:

- eine ausreichend detaillierte Beschreibung der charakteristischen Merkmale
- eine Entscheidungsprozedur, mit der zielstrebig die Menge der im Korpus archivierten Bilder auf einige wenige eingeengt werden kann
- ein Verfahren, das die menschlichen Fähigkeiten des Dialogführenden einerseits und die Fähigkeiten eines Computers andererseits so aufeinander abstimmt, daß sie sich ergänzen
- Aussagen darüber, wie sich ein solches System im praktischen Einsatz verhalten wird

Die Merkmalsbeschreibung

In [1] wird ein 21-dimensionaler Merkmalsvektor benützt mit 5 Unterteilungen in jeder Dimension. Man hat also eine Beschreibungsvielfalt

$$5^{21} \approx 10^{13} \quad \ldots\ldots 10^{14}$$

Im einzelnen gibt es die Merkmalsklassen Haar, Stirn, Backen, Augenbrauen, Augen, Nase, Mund, Kinn, Ohren mit jeweils Unterklassen. Die Unterklassen beim Auge sind z.B. Augenöffnung, Augenabstand, Augentönung. In jeder Unterklasse sind dann 5 Abstufungen.

Die Rangbemessung

Die Auswahl eines gesuchten Bildes kann man sich in mehreren Schritten in der Weise vorstellen, daß der Mensch am dialogführenden Endgerät für ein Merkmal die nach seiner Meinung zutreffende Stufung der Merkmalsbeschreibung angibt und das Archiv durch den Rechner auf möglichst gute Übereinstimmung durchsucht wird. Man braucht hierfür ein quantitatives Maß. Dies erhält man, indem man von der Merkmalszuordnung die beim Archivieren jedem Bild offiziell beigegeben wird, zahlenmäßig die Abweichung feststellt. Beim Auswahlprozess sind dann kleine Abweichungen gut zu bewerten und große Abweichungen stark negativ; dazwischen ist ein zweckmäßiger Übergang zu wählen.

In dem vorliegenden System wurde dies dadurch erreicht, daß man die Absolutbeträge der Merkmalsabweichungen, die man bei den einzelnen Auswahlschritten feststellt, aufsummiert, also $\Sigma|\Delta|$ bildet und damit dann die sogenannte Rangbemessungszahl

$$\exp\ (-\Sigma|\Delta|)$$

berechnet. Dieser Ausdruck hat bei keiner Abweichung, also $\Sigma|\Delta| = 0$, den Wert 1 und geht mit zunehmender Abweichung asymtotisch gegen 0. Ordnet man nun alle archivierten Bilder nach abfallendem Wert dieser Größe, so werden sich die Bilder mit wenig Abweichungen in den ersten Rängen befinden. Ein Bild wird ferner umso eindeutiger bestimmt sein, wenn es sich nicht nur auf dem ersten Rang befindet, sondern wenn das Verhältnis seiner Rangbemessungszahl zur Bemessungszahl der Bilder auf den nächst niederen Rängen auch stark von 1 abweicht.

Die Rolle von Mensch und Computer beim Auswahlprozess

Auf Grund ihrer Erfahrung sind Menschen in hohem Grade geeignet, auffallende Merkmale, z.B. abstehende Ohren, schnell festzustellen. Es ist auch zweckmäßig, auffallende Merkmale zuerst für den Auswahlprozess einzusetzen. Dabei ist die Wahrscheinlichkeit groß, daß man sich auf die in Frage kommenden Fälle zielstrebig einengt. Wie hat man jedoch vorzugehen, wenn auffallende Merkmale nicht vorhanden sind oder zu einer Identifizierung allein nicht ausreichen.

Dann kann man statistische Eigenschaften des gesamten Kollektives ausnützen. Man muß ein Merkmal wählen, das eine möglichst gleichmäßige Häufigkeitsverteilung über den Wertbereich hat. Dann hat man ohne Zusatzinformation - und es war ja vorausgesetzt, daß auffallende Merkmale nicht mehr da sind - die bestmögliche Chance, das Kollektiv weiter einzuengen. Statistische Zusammenhänge aber kann ein Computer leicht überprüfen.

Ein gutes System hat man sich also in der Weise vorzustellen, daß der beurteilende Mensch zunächst über die auffallenden Merkmale suchen läßt und anschließend Merkmale benützt, die mit Computerunterstützung nach statistischen Gesichtspunkten gewählt werden.

Anzahl der Suchschritte

In einem Kollektiv mit N Elementen benötigt man nach Aussagen der Informationstheorie eine Informationsmenge von

$$ld\ N \quad Bit$$

Bei einer 5-stufigen Merkmalsbewertung hat man pro Schritt einen Informationsgewinn von

$$ld\ 5 = 2.3\ Bit/Schritt$$

Daraus läßt sich die Zahl der Suchschritte, die im günstigsten Falle nötig sind, berechnen zu

$$\frac{ld\ N}{ld\ 5} \quad Schritte$$

Für das in Frage kommende System mit einem Archiv von 255 Bildern folgt daraus eine optimale Schrittzahl von 3,4

Das Suchverfahren im praktischen Einsatz

- Übereinstimmung der Merkmalseinstufung mit der offiziellen Beschreibung.

  Die Abweichung von der offiziellen Mermalseinstufung war in 95 % der Fälle höchstens eine Stufe und weniger. Da die Testpersonen sich mit der Arbeitsweise des Systems vertraut machen konnten, sonst aber keine besonderen Fähigkeiten vorausgesetzt wurden, kann man schließen, daß Menschen für derartige Klassifizierungstätigkeiten normalerweise gut geeignet sind.

- Leistungsfähigkeit des Systems

  Es wurden hierüber 3 Testreihen geführt:

  (a) automatische Auswahl des nächsten Merkmales

  (b) Auswahl des nächsten Merkmals durch den Menschen

  (c) Einsatz der beiden Verfahren je nach Zweckmäßigkeit.

Die besten Ergebnisse wurden mit der gemischten Methode erreicht. Etwas schlechter war das Ergebnis, wenn die Reihenfolge vom Menschen allein gewählt wurde und nur halb so gut, wenn die Auswahl vom Computer allein gewählt wurde. Als Maßstab diente dabei der Rang, den das zu bestimmende Bild am Ende der Auswahlsitzung hatte.
Das Ergebis unterstreicht den großen Einfluß hervorstechender Merkmale, nach denen sich ja der Mensch richten kann und dann auch instinktiv richtet. Die nähere Analyse der Ergebnisse zeigt aber auch, daß verhältnismäßig einfache Verfahren zur automatischen Merkmalsauswahl sehr praktikable Ergebnisse liefern. In mehr als 2/3 der Testfälle hatte das gesuchte Bild einen Rang von besser als zehn.

- Anzahl der Suchschritte

  Weiter oben ist für das untersuchte System als optimale Schrittzahl 3.4 abgeleitet worden. Im Testbetrieb waren dagegen meist 8 bis 1o Schritte nötig. Mit Simulationen wurde geklärt, ob der optimale Wert überhaupt erreichbar ist und wenn ja, ob die Reihenfolge oder Abweichungen in der Bewertung der Merkmale für die Verlängerung der Schrittzahl wesentlich sind. Die Simulationsergebnisse zeigen, daß Abweichungen in der Bewertung das Ergebnis stärker beeinflussen als die Reihenfolge, in der die Merkmale benützt werden. Werden diese beiden Einflüsse in ihrem günstigsten Verhalten simuliert, erhält man auch die optimale Schrittzahl. Daraus ist ebenfalls zu schließen, daß man bezüglich eines Auswahlalgorithmus nicht allzu aufwendig zu sein braucht.

- Erweiterung des Archivs

  Im praktischen Einsatz muß man Archive beherrschen, die wesentlich größer als 255 Elemente haben werden. Wie wirken sich Erweiterungen auf die Betriebskosten aus? Dies ist die wichtigste Frage, da die Realisierungsmöglichkeit außer Zweifel steht. Da aus Systemüberlegungen die Gesamtbetriebskosten proportional sind zum Produkt aus Bildkorpusgröße und Logarithmus der Bildkorpusgröße, kann man schließen, daß bei heutigen Preis-Leistungsverhältnissen von Computern Archiv-Größen von 5ooo bis 1o ooo Elementen leicht handhabbar sind.

Die natürliche Sprache und der Computer

Die Sprache hat im menschlichen Kommunikationsprozess wohl die größte Bedeutung. Die Erschließung ihres Formenreichtums für die Computer-unterstützte Informationsverarbeitung und Wiedergewinnung hätte große Vorteile, unter anderem

- Man brauchte sich keine neuen Formalismen im Umgang mit Informationssystemen anzueignen
- Der Mensch ist gewohnt, die für ihn wichtigen Dinge mit dem Instrument Sprache zu beschreiben

- Wir könnten beginnen auch solche Informationsverknüpfungen zu automatisieren, die uns heute noch unmöglich sind

Für die Realisierung eines solchen Unterfangens sind zwei Dinge nötig

- Sätze unserer Sprache müssen von einem Automaten richtig in ihre syntaktischen Elemente zerlegt werden können
- Es sind computergerechte Formen für die Darstellung der Semantik, die über die z.Zt. gebräuchlichen hinausgehen,zu schaffen

Stand der Syntaxerkennung

Die automatische Syntaxerkennung ist der Teil der Computertheorie, der sehr weit durchleuchtet ist. Fußend auf den Erkenntnissen, die mit Programmiersprachen gewonnen wurden, und auf neueren linguistischen Forschungen versucht man an verschiedenen Stellen auch den Formenreichtum von natürlichen Sprachen automatisch handhabbar zu machen.

Jede automatische Sprachanalyse benötigt ein entsprechendes Vokabular und eine minder oder mehr große Anzahl von Produktionsregeln. Bei den natürlichen Sprachen sind nicht nur die Beherrschung der Vielzahl von Produktionsregeln Gegenstand eifriger Forschungen sondern auch die Beherrschung des umfangreichen Volabulars. Der größere Formenreichtum der deutschen Spache gegenüber dem Englischen wird die Beherrschung der Satzproduktion sicher erschweren, er wird es jedoch wahrscheinlich ermöglichen, das Wörterbuchproblem besser als im englischen in den Griff zu bekommen. Dieser Punkt ist gegenwärtig an verschiedenen Stellen Gegenstand der Forschung.

Stand der Semantik-Darstellung

In der Semantik können wir mit einer Definition der Information, die - wie in der klassischen Informationstheorie - auf der Nachrichten-Länge basiert, nichts anfangen. Die zahlreiche Literatur über Informationstheorie kann darüber nicht hinweg täuschen.

Ein erster brauchbarer Ansatz ist die Verwendung von Deskriptoren, die in einem Thesaurus zusammengefaßt sind. Deskriptorwerte sagen dem Fachmann sehr viel; sie haben jedoch die Einschränkung, daß sie durch Weglassen des Kontextes einen Informationsverlust erlitten haben. Für allgemeinere automatisierte Analysen sind sie daher nur bedingt einsatzfähig.

Eine Form die sicherlich aussagefähiger als Deskriptoren ist, wird im Prädikatenkalkül verwendet. Sie basiert auf dem Funktionsbegriff der Mathematik in der Art

$$F(A) = B$$

wobei von einer Grundmenge A ausgegangen wird, die mit Hilfe von F auf eine Menge von Attributen B abgebildet wird. Beispiele hierfür mögen sein

OP (TUMOR)=MAGEN	Durch Operation wird (wurde) ein Tumor aus dem Magen entfernt
LAGE(SPLITTER)=AUGE	Im Auge befindet (befand) sich ein Splitter

Mit dieser Darstellungsform wurde ein computerunterstütztes System für das operative Berichtswesen als Demonstrationsmodell aufgebaut, in das die Operationsberichte in Klartext maschinenlesbar eingegeben werden [2] . Das System hat einen automatischen Syntaxanalysator, der die eingelesenen Informationen in 2o unterschiedliche Funktionen zerlegt. In dieser Form werden sie dann, versehen mit Hinweisen, wo die Information herstammt, als sogenannte Kernsätze abgespeichert.

Diskussionen mit einem der Urheber ergaben folgende Bewertung

- Die Darstellung als Kernsätze wird als großer Fortschritt gegen bisherigen semantischen Formen angesehen

- Das System wurde letztlich nicht zu einem operablen System zu Ende entwickelt, weil es nicht gelang, die Entwicklungsmannschaft zusammen zu halten.

Die gezeigte Funktionsdarstellung erlaubt nur einfache Aussagen semantisch zu beschreiben und, wie im Prädikatenkalkül mit bestimmten Operatoren miteinander zu verbinden wie "und", "oder","nicht", "äquivalent". Es fehlt jedoch die Darstellungsmöglichkeit für Relationen von Relationen. Der Satz, daß der gestürzte Patient nach der Amputation des Beines starb, benötigt kompliziertere Darstellungsformen als die obige, will man keinen wesentlichen Informationsverlust haben. Eine mögliche Form könnte sein [3]

NACH(AMP(BEIN), STERBEN(STÜRZEN(PATIENT)))

Mit ähnlichen Darstellungsformen wird an verschiedenen Stellen gearbeitet. Ihr großer Vorteil ist darin zu sehen, daß derartige Aussagen mit anderen auf Widerspruchsfreiheit automatisch überprüft werden können. Überprüfbarkeit auf Widerspruchsfreiheit spielt für die Weiterentwicklung von Programmiersprachen eine große Rolle. Es ist anzunehmen, daß die hier vorliegende Thematik aus den bei den Programmiersprachen gewonnenen Erkenntnissen neue Impulse bekommt. Der derzeitige Nachteil ist, daß der Rechenaufwand für den praktischen Einsatz noch prohibitiv hoch ist.

Ausblick

Als Techniker hat mein Interesse an diesem Arbeitskreis zwei Zielsetzungen, nämlich herauszufinden, welche Bedürfnisse die Medizin hinsichtlich ihres Vokabulars und hinsichtlich der semantischen Struktur hat.

Hinsichtlich des Vokabulars liegen sehr große Vorleistungen vor.

Hinsichtlich der semantischen Struktur stellt sich mir die Frage, welche Anforderungen gegenwärtig und zukünftig zu stellen sind und welche Mittel wir hierzu benötigen. Das zur Zeit vorherrschende Thesaurusprinzip paßt verhältnismäßig gut in die Möglichkeiten der heutigen Datenbanken. Damit kann man viele aktuelle Probleme des medizinischen Berichtswesens mittels Computerunterstützung modernisieren. Man ermöglicht aber nur unvollkommen, kompliziertere Fragestellungen oder Fragestellungen über medizinische Fachdisziplinen hinweg mit automatisierter Unterstützung zu analysieren. Hierfür werden Hilfsmittel benötigt, die über die Zielsetzung der heutigen Klartextanalyse hinausgehen.

Literatur

[1] Goldstein, Harman, Lesk

Man-Machine. Interaction in Human Face Indentification

The Bell System Technical Journal
Vo 51, No 2, Febr 1972

[2] Shapiro, Stermole

ACORN. An Automated Natural Language Question-Answering System for Surgical Reports

Computers and Automation
Febr 1971

[3] Biss, Chien, Stahl

R2 - A Naural Language Question-Answering System

University of Illinois - Urbana, Illinois
Report 5 -5oo Jan 1971

Informationstheoretische Aspekte medizinischer Routine-Befundmitteilungen

P. Röttger

Senckenbergsches Zentrum der Pathologie

Universität Frankfurt

Bei Erhebung eines medizinischen Befundes wird unterschieden zwischen einem Routinebefund mit dem eng definierten Kommunikationsbedarf eines Einzelfalles und dem auch außerhalb der Routine bedeutsamen Befund von allgemeinem Interesse. Beide Vorgänge unterscheiden sich nicht bezüglich des mitgeteilten Substrates, des medizinischen Sachverhaltes. Die Art und Weise, in der der Befund von allgemeinem Interesse einem nicht näher definierten Kreis von Empfängern mitgeteilt wird (medizinische Befundpublikationen) unterscheidet sich grundlegend von der Routinebefundmitteilung. Dies betrifft nicht zuletzt auch das verwendete Kommunikationsmedium, den Befundtext.

In unserer Betrachtung wollen wir uns 1. mit der Definition des medizinischen Sachverhaltes, 2. mit dem Informationsfluß im Routinebetrieb, seinem formalen Ablauf und seinem Kommunikationsziel einschließlich der speziellen Ansatzmöglichkeiten der Textweiterverarbeitung sowie schließlich 3. mit den sich daraus ergebenden Konsequenzen für den Routinebetrieb und für die Konzeption von Informationssystemen befassen.

1. Das Substrat des Kommunikationsvorganges

Das Substrat des Kommunikationsvorganges im medizinischen Routinebetrieb ist die Beobachtung eines Sachverhaltes (SV). Hierunter verstehen wir bei Untersuchung eines Patienten die Beobachtung einer Abweichung von der strukturellen und funktionellen Norm. Die Feststellung eines derartigen Sachverhaltes schließt das qualitative und das quantitative Ausmaß sowohl des pathologischen Befundes als auch der physiologischen Restfunktion und der normal-anatomischen Reststruktur ein. Bei einem Untersucher mit umfangreicher medizinischer Erfahrung addieren sich die beobachteten Sachverhalte zu übergeordneten Begriffsfeldern. Das Patientenkollektiv, das er persönlich überblickt, vermittelt ihm bezüglich der Erkenntnis eines Sachverhaltes eine gewisse Erfahrungsbilanz, von der er bei der jeweiligen neuen Bewertung einer Einzelsituation ausgehen kann.

Wir haben versucht, diese Vorstellungen bei der Auswertung von Patientenkollektiven weiter zu definieren (LÜHR, 1972). Dabei haben wir die medizinischen Sachverhalte unterteilt in Sachverhalte I. Ordnung, Sachverhalte II. Ordnung und Sachverhalte III. Ordnung.

Unter Sachverhalte I. Ordnung haben wir die Summe aller Beobachtungen pathologischer Befunde in einem festgelegten Patientenkollektiv verstanden, unter einem Sachverhalt II. Ordnung die Summe aller Befunde in Zusammenhang mit einer übergeordneten Lokalisation (z.B. im Organ "Leber"). Als Sachverhalt III. Ordnung verstand sich ein Befund bei einem Einzelprobanden, der sich von anderen Befunden bei dem betreffenden Probanden inhaltlich abgrenzen läßt.

Im Rahmen der Mitteilungen pathologisch-anatomischer Befunde in einem Autopsie-Bericht wird diese Einzelbeobachtung eines Sachverhaltes als ein sog. Einzeldiagnosesatz niedergelegt. Es besteht also das Prinzip, den Einzelsachverhalt in einem logisch zusammengehörigen Diagnosesatz mitzuteilen. Diese Basis ist sowohl für die Analyse der Routinekommunikation als auch für die Konzeption von Auswertungssystemen bedeutsam, so weit sie sich mit den Informationen befassen, die innerhalb der ärztlichen Routinetätigkeit anfallen.

2. Der Informationsfluß

Wir gliedern diese Darstellung in eine Beschreibung des formalen Ablaufs des Kommunikationsvorganges sowie in eine Darstellung des angestrebten Kommunikationszieles.

2.1 Der formale Ablauf

In Abb. 1 wird schematisch veranschaulicht, welchen Weg die Mitteilung eines Sachverhaltes in der zwischenärztlichen Kommunikation geht. Voraussetzung für diese Betrachtung sei die Mitteilung eines einseitig zugänglichen Sachverhaltes, z.B. die Mitteilung über eine histologische Befundung eines zur Untersuchung an einen Pathologen eingesandten Organs (Biopsiebefund).

Der Untersucher A (Pathologe) erfaßt den vorliegenden Sachverhalt sowohl makro- als auch mikroskopisch und hat aufgrund dieser Untersuchung eine Reihe von Gedankeninhalten. Der Sachverhalt (SV) bildet sich also gedanklich im Untersucher A ab, was wir mit dem Symbol

$$SV, A \longrightarrow A\ (SV)$$

bezeichnen.

Um diesen Gedankeninhalt seinem Kommunikationspartner, dem Arzt B (Kliniker) mitteilen zu können, muß der Untersucher A seinen Gedankeninhalt zum Sachverhalt in eine bestimmte Form, d.h. in Worte überführen, ein Vorgang, der im weitesten Sinne als "Encodierung" zu verstehen ist. Als Resultat der Encodierung liegt im Normalfall die Beobachtung des Sachverhaltes in schriftlicher Form vor. Die Informationsmenge in diesem Textdokument kennzeichnen wir mit dem Symbol

$$A\ (SV)\ Enc$$

Wird dieses Dokument dem Zweck der Kommunikation entsprechend dem Untersucher B zugeleitet, so muß dieser es gedanklich erfassen. Bei diesem Vorgang werden Worte bzw. ein Wort-Code wieder in Gedankeninhalte überführt, er ist also als "Decodierung" zu definieren. Als Resultat dieser Decodierung bildet sich beim Untersucher B ein bestimmter Gedankeninhalt zu der aufgenommenen Mitteilung "A (SV) Enc". Diese gedankliche Abbildung des veranschaulichten Sachverhaltes beim Untersucher B kann also definiert werden als

$$B\ \left[A\ (SV)\ Enc\right] .$$

Als Abschluß des Kommunikationsprozesses bildet sich der

beobachtete Sachverhalt (SV) in beiden Untersuchern in Gestalt bestimmter Gedankeninhalte ab, das Kommunikationsziel ist erreicht.

2.2 Das Kommunikationsziel (s.Abb. 2a)

In Abb. 2a ist das Ziel der zwischenärztlichen Kommunikation symbolisch veranschaulicht: Zwischen den Gedankeninhalten des Untersuchers A zum Sachverhalt SV - "A (SV)" - und dem Gedankeninhalt des Untersuchers B zum gleichen Sachverhalt (der ihm aufgrund der Übermittlung des Dokumentes mit den encodierten Gedanken von A möglich geworden ist) - "B [A (SV) Enc] " - ist eine möglichst große Übereinstimmung herzustellen. Die Grenzen der Kommunikation bringen es mit sich, daß eine solche Übereinstimmung nur selten fast vollständig erreicht werden kann. Meist kommt nur eine relativ kleine Schnittmenge zwischen den entsprechenden Gedankeninhalten beider Untersucher zustande. Diese Schnittmenge ist jedoch u.U. von entscheidender Bedeutung für das Schicksal des Patienten. Die weitere Diagnostik und ggf. die Therapie hängt von diesem Kommunikationserfolg ab, den wir symbolisch mit

$$A,B\ (SV)$$

kennzeichnen.

Die Qualität des Kommunikationserfolges wird von vier Faktoren bestimmt, die wir in diagnostische Faktoren und Kommunikationsfaktoren unterteilen können.

2.2.1 Die diagnostischen Faktoren (s. Abb. 2b)

Nicht jeder bei einem Patienten erhobene medizinische Sachverhalt wird in gleicher Weise von dem Untersucher A erkannt und in Form von Gedankeninhalten in ihm abgebildet werden können. Die Qualität der Untersuchung ist abhängig vom allgemeinen Erkenntnisstand der betreffenden Spezialwissenschaft. Der Untersucher A partizipiert an diesem Gesamtwissen seines Faches in einem unterschiedlich großen Anteil. Von dem Ausmaß dieses Anteils hängt es ab, in welchem Umfang der Untersucher A den medizinischen Sachverhalt zu erfassen vermag. Die Beziehungen zwischen Realität (SV) und Informationsgewinnung sind in Abb. 3 sybolisch dargestellt.

Das Zeichen "$\bar{A}$ (SV)" gibt an, welcher Begriffs- bzw. Gedankeninhalt über SV dem gesamten allgemeinen Erkenntnisstand nach maximal möglich ist. Diese Informationsmenge ist naturgemäß größer als die Menge "A (SV)", die in ihrem Ausmaß durch die idividuelle Kapazität des Untersuchers bestimmt wird. Die Qualität der Einzeluntersuchung, d.h. die Menge an Information, die in der Position "A (SV)" anfällt, ist also einmal abhängig von der Differenz

$$SV - \bar{A}\ (SV),$$

d.h. von der Qualität der betreffenden Fachwissen-

schaft i.e. von der Kapazität aller möglichen Untersucher, zum anderen ist sie abhängig von der Differenz

$$\bar{A}\ (SV) - A\ (SV),$$

d.h. abhängig von der Qualität der Einzeluntersuchung i.e. der Kapazität des zufälligen Untersuchers A.

Beide Momente bestimmen die Kapazität der jeweiligen Diagnostik. Es handelt sich mithin um die Faktoren, die den Kommunikationserfolg von der Basis her (von "A (SV)" aus) mitbestimmen.

2.2.2 Kommunikationsfaktoren

In dem in Abb. 1 dargestellten Schema ist ersichtlich, daß im Rahmen des Informationsflusses von A nach B der Untersucher A seine Gedankeninhalte zu encodieren, der Untersucher B diese encodierten Gedankeninhalte wieder zu decodieren hat. Bei diesen Vorgängen ist mit Informationsverlusten zu rechnen, also damit, daß Informationen verlorengehen bzw. unvollständig wiedergegeben werden können.

In Abb. 2c wird der Vorgang des Encodierens beim Untersucher A veranschaulicht: Es wird gezeigt, daß entsprechend der Encodierungsqualität von A nur ein Teil seiner Gedankeninhalte zu SV, also nur ein Teil der Informationsmenge A (SV) weitergegeben werden kann.

Beim Untersucher B ist es von seinem Verständnis für die encodierten Gedankeninhalte von A abhängig, in welchem Ausmaß er Informationen über SV aufzunehmen vermag. Diese Decodierungsqualität verlangt ein gewisses Ausmaß an Einsichten in das Spezialfach des Untersuchers A. Sie kann durch Vorkenntnisse von B über SV (z.B. Röntgenbefund, Labordaten, die ihn veranlaßt haben, sich mit einer bestimmten Fragestellung zu dem vermuteten SV an A zu wenden) verbessert sein.

Diese Zusatzkomponente im "normalen Kommunikationsprozeß haben wir in unser Schema nicht einbezogen. Des besseren Verständnisses wegen sind wir von der "abnormen" Situation ausgegangen, daß der von A aufzunehmende SV einseitig zugänglich wäre.

2.3 Der Informationsgehalt der verschiedenen Stufen des Kommunikationsprozesses

Wenn wir das maximal mögliche Ausmaß an Informationen über SV, d.h. den Sachverhalt an der Realität als Ausgangsbasis nehmen, so ergeben sich bis zum Kommunikationsziel "A,B (SV)", dem gemeinsamen Gedankeninhalt beider Untersucher zu SV folgende Informationsschwellen:

1) $SV > \bar{A}\ (SV)$.

   Der Sachverhalt als Realität kann selbst bei Anwendung aller Kenntnisse eines Spezialfaches nur teilweise als Gedankeninhalt abgebildet werden.

2) $\overline{A}$ (SV) $>$ A (SV).

Der Untersucher A ist bei vielen Spezialproblemen nicht in der Lage, auf das Gesamtwissen seines Faches zurückzugreifen. Seine Kapazität ist von seinem Ausbildungsstand und seinen technischen Möglichkeiten sowie von einem individuellen Erfahrungshintergrund (Sachverhalte I. und II. Ordnung) abhängig.

3) A (SV) $>$ A (SV) Enc

Der Untersucher A kann seine Gedankeninhalte zu SV nur in begrenztem Umfange in Worte überführen. Zwar spielt sich die gedankliche Auseinandersetzung vorwiegend über Wortassoziationen ab, jedoch sind diese freien Wortassoziationen von A über SV nicht notwendig identisch mit der encodierten Darstellung von SV, z.B. hat A zu berücksichtigen, daß der Empfänger B die Probleme des Spezialfaches von A nur in begrenztem Umfange erfassen kann.

4) A (SV) Enc $>$ B A (SV) Enc

Die unterschiedliche "Spezialfachkenntnis" bestimmt auch den Umfang, in welchem B die ihm übermittelte Mitteilung von A zu decodieren, also aus der encodierten Form der Informationen wieder Gedankeninhalte zu bilden imstande ist. Je nach Art des Einzelsachverhaltes wird dies in einem unterschiedlichen Ausmaß der Fall sein. Ein weiterer Informationsverlust ist hier im Regelfalle zu erwarten.

5) B [A (SV) Enc] $>$ A,B (SV).

Da keine völlige Übereinstimmung in den Gedankeninhalten von A und B bestehen kann, ist die Informationsmenge, die wir als gemeinsame Gedankeninhalte beider Unternehmer zu SV bezeichnen, kleiner als die Gedankeninhalte von B über SV. Zwischen "SV" und "A,B (SV)" besteht also eine erhebliche Differenz (zwischen den möglichen und den tatsächlich erfaßten Informationen). Ziel der medizinischen Kommunikation muß es sein, diese Differenz so klein wie möglich zu halten, Ziel einer befundbezogenen - also auf die Analyse von SV ausgerichteten - Datenverarbeitung ,uß es sein, den Zugang zum Informationsfluß zu finden, der der maximal faßbaren Informationsmenge zu SV, nämlich "$\overline{A}$ (SV)" am nächsten kommt.

2.4 Die Ansatzpunkte der Datenverarbeitung (s.Abb.3)

In Abb. 3 haben wir das Schema von Abb. 1 nochmals wiederholt und um eine "Seitenkette" des Informationsflusses ergänzt, die sich bei manueller Verschlüsselung des Routinedokumentes ergeben würde. Diese Seitenkette setzt dort an, wo auch die maschinelle Auswertung eines Routinedokumentes angreifen muß, nämlich an den encodierten Gedankeninhalten von A über SV, also an der im Regelfalle schriftlich niedergelegten Form des beobachteten Sachverhaltes.

2.4.1 Die vollautomatische Textverarbeitung

Bei der Besprechung der theoretischen Voraussetzungen des AGK-Thesaurus werde ich auf die Detailprobleme in dieser

Position im einzelnen zurückkommen. Aus der speziellen Kommunikationsintention - der aktuellen Information von B über SV unter Voraussetzung bestimmter medizinischer Grundkenntnisse - läßt sich jedoch ohne weitere Erklärung ersehen, daß das für diesen Zweck konzipierte Dokument "A (SV) Enc" in seiner Form einer automatischen Datenverarbeitung nicht oder nur begrenzt zugänglich sein wird (Röttger u. Mitarb., 1969).

Es muß sich also an die Routine-Dokumentation ein Bearbeitunsvorgang anschließen, der 1. ein gewisses Äquivalent für die durch B vorzunehmende Decodierung beinhaltet und der 2. die beschränkten Einsichten von B in die Probleme des Spezialfaches von A bei dem automatischen Auswertungssystem zu kompensieren vermag. Das Auswertungssystem muß sich also nach dem Kenntnisstand des Spezialfaches von A richten. Die Erfassung des Sachverhaltes sollte durch Einbringung von Zusatzinformationen in eine möglichst enge Beziehung zu dem derzeitigen aktuellen Gesamtkenntnisstand eines Spezialfaches gebracht werden. Das Ergebnis der automatischen Verarbeitung soll also weitmöglichst in Richtung auf die Übertragung von

$$A\ (SV)\ Enc \rightarrow \bar{A}\ (SV)$$

ausgerichtet werden

2.4.2 Das manuelle Verschlüsseln (s. Abb. 4)

In Abb. 4 ist die akzessorische Seitenkette des Informationsflusses nochmals gesondert dargestellt. Dieser zusätzliche Ablauf spielt sich innerhalb des Spezialfaches von A ab. Seine encodierten Gedankeninhalte zu SV werden einem weiteren Kollegen übermittelt, der sie decodiert und diese Gedankeninhalte in die Symbolik eines Schlüsselsystems überträgt. Diesen Vorgang, der nach der Decodierung abläuft, bezeichnen wir als Metacodierung. Die metacodierten Informationen liegen systematisch geordnet in einer Form vor, die eine automatische Auswertung zuläßt. Bis zu diesem Endergebnis - C [A (SV) Enc] Metac - sind zwei zusätzliche Stufen eingeschaltet.

1. A (SV) Enc $>$ C [A (SV) Enc],.

   Die Kommunikation innerhalb eines Spezialfaches (die Kommunikation zwischen A und C) steht unter einer anderen Problematik als die Kommunikation zwischen Ärzten aus verschiedenen Spezialfächern. Der Informationsverlust in dieser Position dürfte geringer sein als der zwischen A und B.

2. C [A (SV) Enc] $>$ C [A (SV) Enc] Metac

   Die zweite Informationsschwelle in dieser Seitenkette ist bedeutsamer als die erste. Ihre Höhe wird durch drei weitere Faktoren bestimmt:

1) Durch die Kenntnis von C von der speziellen Problematik des jeweiligen SV,
2) durch die Kenntnis von C über die interne Struktur des Schlüsselsystems,
3) durch die Qualität und Aktualität des verwendeten Schlüsselverfahrens.

Zu 1 und 2: Damit die Informationsverluste, die mit der Person von C verbunden sind, nicht zu hoch werden, verbietet sich weitgehend der Einsatz von Anfängern in einem Spezialfach für die manuelle Verschlüsselung. Eingearbeitetes medizinisches Hilfspersonal vermag wertvolle Unterstützung zu leisten, ist jedoch bei schwierigen Problemen nur begrenzt entscheidungsfähig.

Zu 3: Die Kapazität des Verschlüsselungssystems wird begrenzt dadurch, daß es bis zu einem wesentlichen Anteil "benutzerfreundlich" konzipiert werden muß. Zum anderen muß bei der Konzeption eines festen Verschlüsselungssystems auf lange Sicht geplant werden. Das beinhaltet, daß der aktuelle Informationsfluß innerhalb eines Spezialfaches nur in begrenztem Umfang in die Schlüsselkonzeption einfliessen kann. Bezeichnen wir den allgemeinen Kenntnisstand eines Spezialfaches zum Zeitpunkt einer Schlüsselkonzeption (Metacodierungssystem) mit $\overline{M}$, so ist auch die bezüglich jedes Spezialproblemes unterschiedliche Aktualität von $\overline{M}$ ein weiterer Faktor, der die Höhe der Informationsschwelle in dieser Position des Kommunikationsprozesses mitbestimmt.

3. Wechselseitige Beeinflußung von Kommunikation und Dokumentation

Es bedarf keiner Begründung, daß die Belange der aktuellen Kommunikation vorrangig bleiben müssen. Das "eingespielte" Encodierungs- und Decodierungs-System zwischen A und B sollte möglichst wenig verändert, nach Möglichkeit verbessert werden. Deswegen sollen eingeführte Benennungen allein auf Anforderung der Dokumentation nach Möglichkeit nicht geändert werden. Sind Änderungen jedoch unvermeidlich, so sollten sie mit einem Maximum an Information unter allen Kommunikanten eingeführt werden, damit das Ausmaß der gedanklichen Schnittmenge über die mitgeteilten Sachverhalte nicht reduziert wird.
Die Existenz eines Datenverarbeitunssystems sollte den Partizipanten nicht nur theoretisch "bekannt" sein. Das Datenverarbeitungssystem soll soweit benutzerfreundlich strukturiert sein, daß der Zugriff zu den gespeicherten Informationen im Routinebetrieb auch den Kommunikanten (A und B) möglich ist. Der Erfahrungshintergrund des Einzeluntersuchers soll durch die Informationsmöglichkeiten erweitert, seine Routinearbeit damit erleichtert und verbessert werden. Gleiches gilt für das wechselseitige Verständnis zwischen den Spezialfächern. Erst damit kann die Benutzerfreundlichkeit eines Systems nicht nur in einer oft lediglich erhofften Freundlichkeit zu Benutzern (=Anwendungskomfort) zum Ausdruck kommen, sondern auch zu einer Freundlichkeit von Benutzern zu einem Dokumentationssystem führen. Ohne eine derartige wechselseitige positive Einstellung wird sich kein Dokumentationsverfahren innerhalb des medizinischen Routinebetriebes auf die Dauer behaupten können.

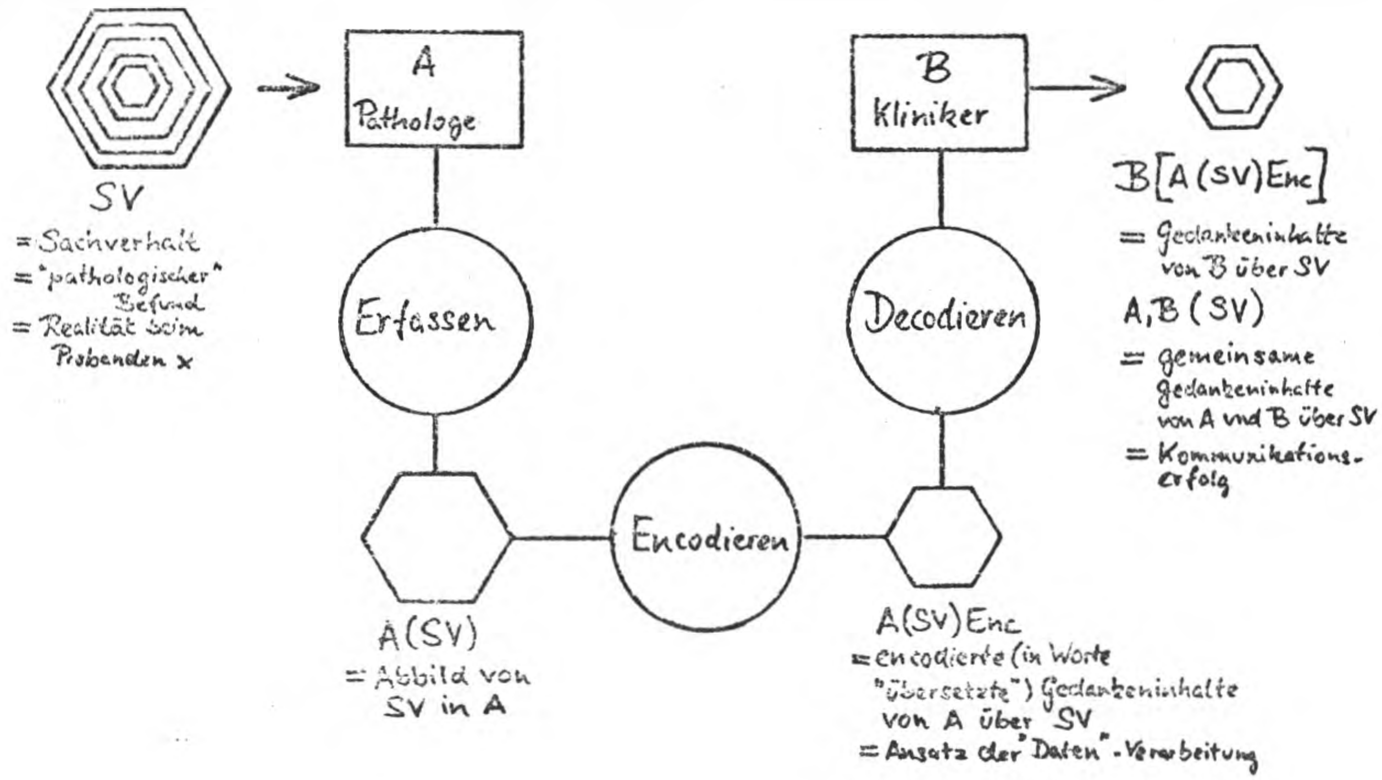

Abb.1: Informationsfluß bei einseitig zugänglichem medizinischen Sachverhalt, z.B. bei histologischer Beurteilung eines Biopsie-Befundes ohne wesentliche klinische Vordiagnostik.

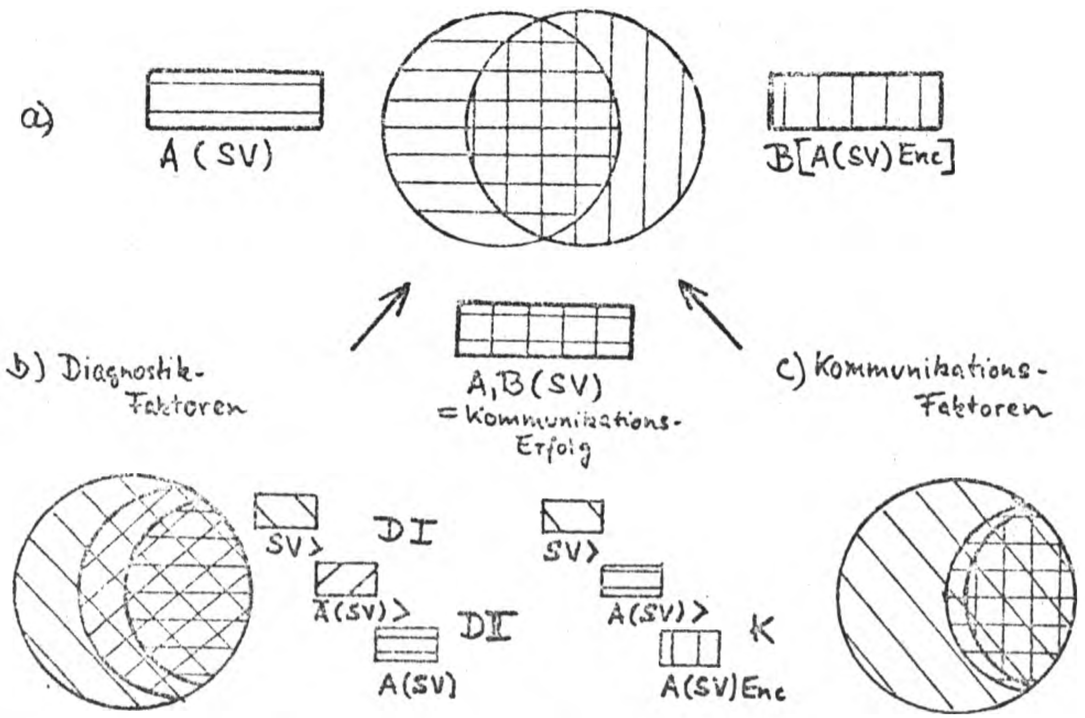

Abb.2: SV = Sachverhalt (i.e. Kommunikations-Substrat)

$\bar{A}$ (SV) = dem gesamten allgemeinen Kenntnisstand eines medizinischen Spezialfaches nach möglicher Gedankeninhalt zu SV

A (SV) = individuell (Untersucher A) maximal möglicher Gedankeninhalt zu SV

DI = Differenz SV - $\bar{A}$ (SV)
= Qualität der Fachwissenschaft von A
= Kapazität aller möglichen Untersucher im gleichen Fachgebiet

DII = Differenz $\bar{A}$ (SV) - A (SV)
= Qualität der Einzeluntersuchung
= Kapazität des Untersuchers A (inclusive "Tagesform")

K = Differenz SV - A(SV)Enc
= Unterschied zwischen der Realität SV zu den encodierten Gedankeninhalten von A über SV, festgelegt durch
1. den Gedankeninhalt A (SV) - s.o. !
2. die Encodierungs-Kapazität von A

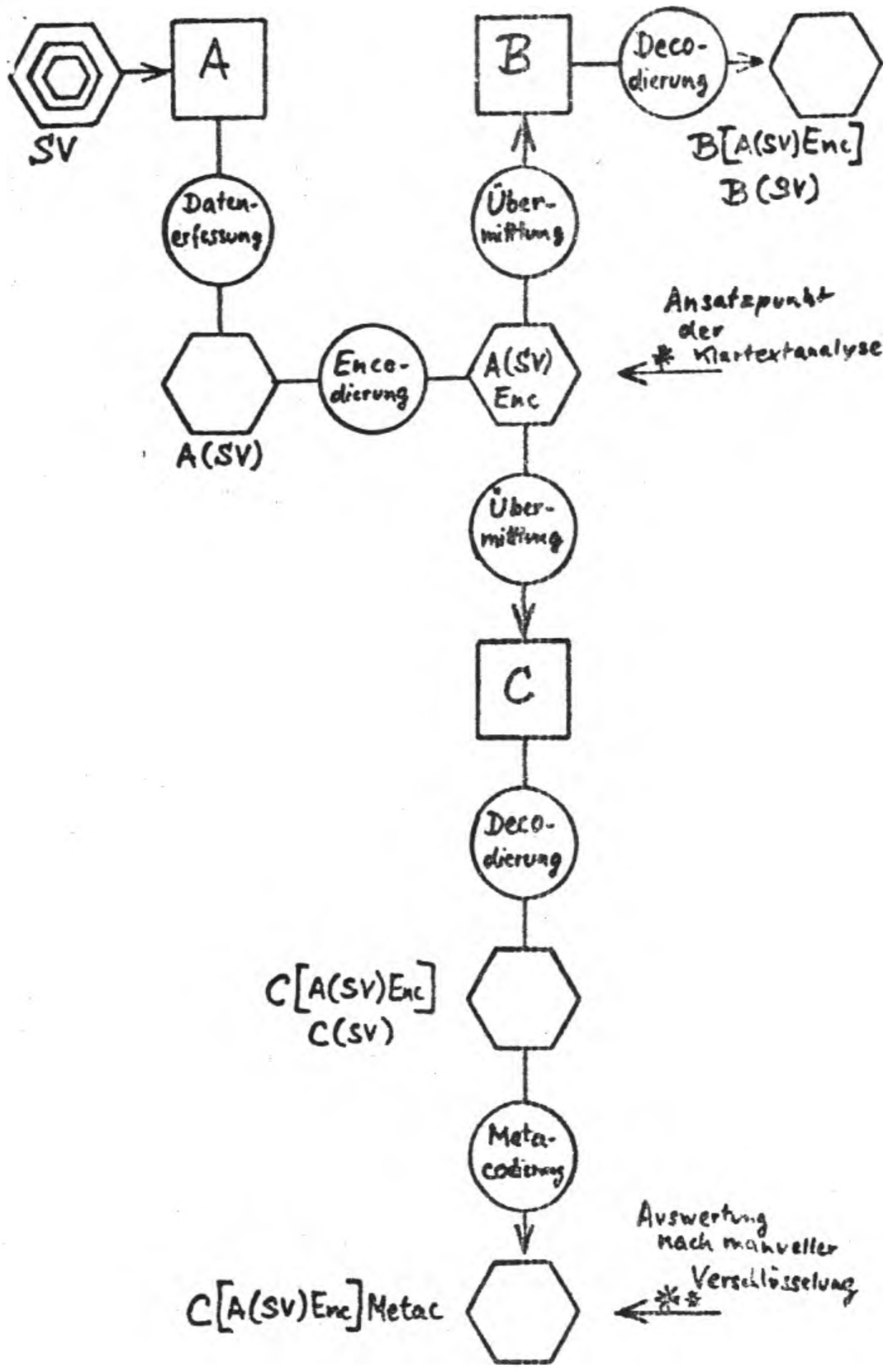

Abb.3: Ansatzpunkte der Klartextanalyse und der Auswertung manuell verschlüsselter Text-Informationen in einem medizinischen Kommunikationsprozess, Beispiel Pathologie-Klinik.
A = datenerhebender Pathologe
B = Kliniker
C = datenverarbeitender Pathologe (Schlüssel-Spezialist)

Encodierung = Übertragung von Gedankeninhalte in Befundtexte

Decodierung = gedankliches ERfassen von Befundtexten

Metacodierung = Übertragung von Befundtexten in ein festliegendes, nicht mehr mit üblichen Methoden (d.h. durch Erfassen von Worten) "lesbares" Schlüsselsystem.

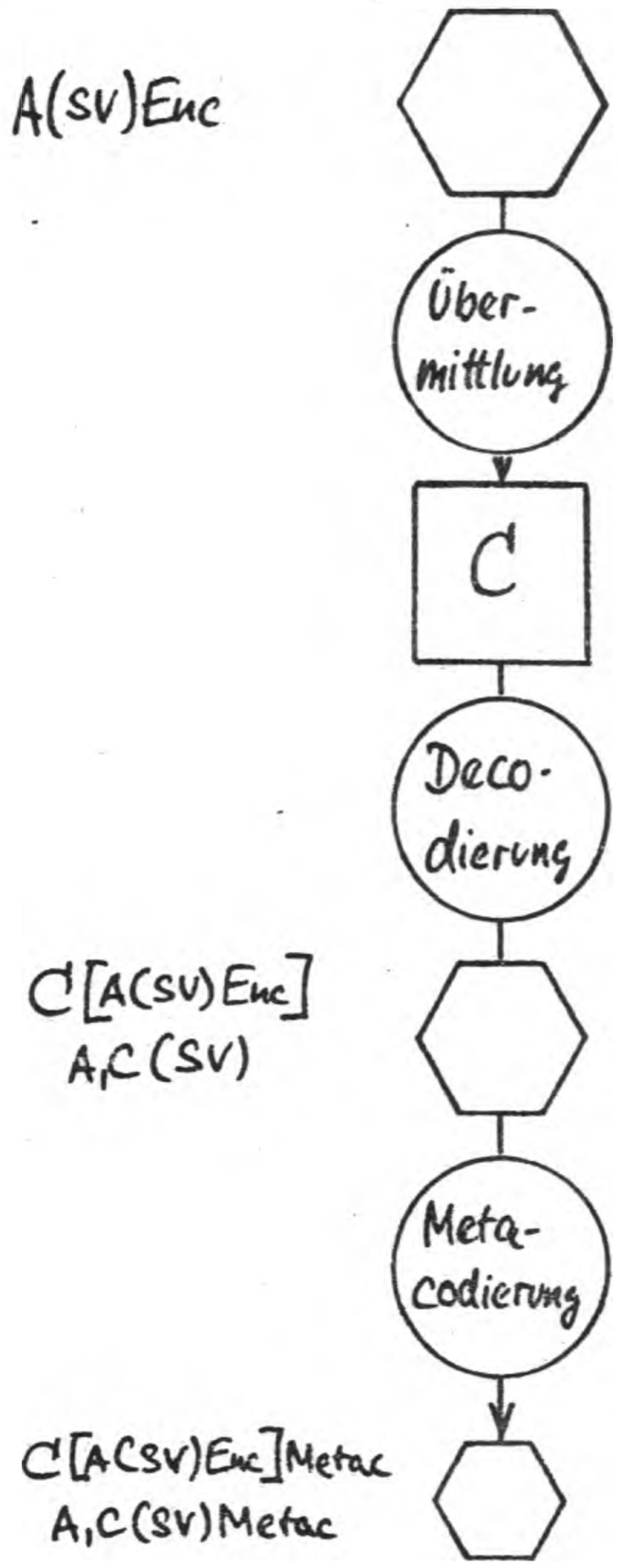

Abb.4: Akzessorische Informationsschwellen bei Auswertung manuell verschlüsselter Daten

I. A(SV)Enc → C (A(SV)Enc)
= Kommunikations-Qualität innerhalb des Spezialfaches von A und C

II. C (A(SV)Enc) → C (A(SV)Enc) Metac
= Schlüssel-Kapazität

Metacode = geschlossenes System, festgelegt durch $\bar{M}$ = Stand des Fachwissens zum Zeitpunkt der Schlüsselerstellung

A,C(SV)Metac = Abbild von SV innerhalb des Schlüsselsystems, bestimmt durch
1. die Qualifikation von A (datenerhebender Pathologe)
2. die Qualifikation von C (datenverarbeitender Pathologe)
3. die Aktualität von $\bar{M}$

Zur Abwandlung von Informationen auf dem Wege
von der Primär- zur Tertiär-Information
in medizinischen Befundtexten.
(Diskussionsbeitrag zu P. Röttger)

B. Leiber

Dokumentations- und Forschungsabteilung für
klinische Nosologie und Semiotik am
Fachbereich Humanmedizin der Universität Frankfurt a.M.

Alle bisherigen Betrachtungen zum weiten Gebiet der "Klartext-analyse" in diesem Arbeitskreis haben noch einen sehr wichtigen Gesichtspunkt gänzlich unberücksichtigt gelassen, den ich hiermit einmal Ihrer ganz besonderen Aufmerksamkeit anempfehlen möchte. Ich halte ihn aus eigener praktischer Erfahrung für so entscheidend wichtig, daß man ihn vielleicht einmal zum Verhandlungsgegenstand einer eigenen Arbeitskreissitzung machen sollte:

Wer Klartext analysiert und verarbeitet, sollte unbedingt auch genau wissen, wie und unter welchen Bedingungen ein medizinischer Klartext entsteht und damit in die Lage versetzt sein, dessen Aussagefähigkeit und Qualität zu beurteilen. Man muß dabei nämlich stets in sein Kalkül die Erfahrung einbeziehen, daß gegenüber der medizinischen Wirklichkeit jeder Zwang, über diese ein Dokument im Klartext niederzuschreiben, meist zu sehr hochgradigen Informationsverlusten führt. Wie jeder Kenner weiß, wird die im unmittelbaren Arzt-Patient-Feld entstehende, vom Patienten direkt ausgehende hochkomplexe, vielfältige, bunte und aktuelle Primärinformation durch die Transformation zur Dokumentenform zu einer stark informationsreduzierten, vereinfachten und oft stark verkürzten Sekundärinformation. Besonders informationsabschwächend in diesem Sinne wirken sich folgende Umstände aus:

1. Wenn dem Arzt der Sachverhalt der Primärinformation eindeutig oder scheinbar zweifelsfrei erscheint, dann reduziert er die dokumentarische Zweitinformation meist drastisch, erspart sich auf diese Weise viele Worte und macht sich in den seltensten Fällen die Mühe, alle wirklich relevanten Daten auch einzeln niederzuschreiben.
2. Wenn bei der Entstehung der Primärinformation dem Beobachter einige zur Beschreibung des Sachverhaltes unbedingt notwendige Termini nicht verfügbar sind, dann gehen die entsprechenden Daten verloren, weil sie nicht ins Dokument eingehen.
3. Es besteht aber auch die Möglichkeit, daß bei der Entstehung der Primärinformation der an sich richtig beobachtende Arzt unrichtige Termini verwendet. In diesem Falle entstehen gravierende Informationsverfälschungen; manchmal entstehen diese auch bewußt oder unbewußt nach Art des sog. "Gymnasium-Post-Effektes".
4. Wenn die Primärinformationen sehr komplexe und multivariable Fakten beinhalten, die sich mit Worten sehr schwer oder garnicht beschreiben lassen, dann reduziert sich die Sekundärinformation oft auf ein Minimum an relevanten Daten.
5. Darüberhinaus gibt es gerade in der klinischen Medizin, sicher auch in der pathologischen Anatomie, zahlreiche Informationen, die der Arzt wohl sieht, erkennt, riecht oder fühlt und selbstverständlich auch für seine unmittelbare Diagnosestellung verwendet, die er aber nicht in Worte umsetzen kann, weil es für sie kein Wort gibt. Diese Primärinformationen kommen somit nicht als Sekundärinformationen in das Dokument und fehlen absolut. Hierher gehörige Informationen betreffen z.B. Form und Bildung eines anomalen Gesichtes, Ohrmuschelanomalien, ferner Geruchseindrücke, Farbnuancen, viele Tastempfindungen, Emotionen und vage Sinneseindrücke u.v.a.m.

Damit sind noch nicht einmal alle informationsverändernden und informationsabschwächenden Einflüsse genannt. Werden nämlich die Sekundärinformationen in Dokumentenform schließlich sprachlich

noch in eine computergerechte Form transformiert, d.h. in eine Tertiärinformation umgewandelt, dann kommt es zwangsläufig zu weiteren, nicht unbeträchtlichen Informationsverlusten. Die praktische Erfahrung zeigt, daß es Situationen gibt, wo das so entstandene Informationsfeeding fast 9o % beträgt. Dies bedeutet, daß der Informationsgehalt der Tertiärinformation sich zum Informationsgehalt der Primärinformation wie 1:9 verhält.

Selbstverständlich gelten die hier nur kurz angedeuteten Umstände ganz allgemein und für die meisten Typen von Dokumenten. Sie sind jedoch in der Medizin, wo von ihnen unter Umständen Diagnose, Therapie, Prognose, Handlungsnotwendigkeiten u.s.w. abhängen, von besonders weitreichender Bedeutung. Vor allem den mit den einschlägigen Fragen einer medizinischen Klartextverarbeitung befaßten Personen müssen diese Probleme weit bewußter werden, als dies bis jetzt der Fall ist. Man muß diesen Personenkreis auch vor dem Trugschluß bewahren, daß eine qualifizierte Klartextanalyse und -verarbeitung mehr sei als nur eine winzige Facette der computergerechten Befunddokumentation und -verwertung.

Definitionen und Voraussetzungen der medizinischen Klartextanalyse

W. Feigl

Pathologisch-Anatomisches Institut der Universität Wien

Als Überleitung vom theoretischen zum speziellen Teil werden die momentan als wesentlich erscheinenden Definitionen und Voraussetzungen der Klartextanalyse gebracht.

Definition der medizinischen Klartextanalyse - was versteht man darunter allgemein und was verbindet diese mit der Klartextanalyse im weiteren Sinn, also dem maschinellen Auswerten beliebiger Texte und

Voraussetzungen - was unterscheidet sie davon und welche Einschränkungen sind zu machen.

Entwicklung und Notwendigkeit der Klartextanalyse

Der medizinische Klartextanalyse kommt immer mehr Bedeutung innerhalb der allgemeinen Klartextanalyse zu. Der sprunghafte Aufstieg verschiedener Systeme und das große Interesse an diesem Gebiet läßt sich am ehesten durch einen kurzen Rückblick über die Entwicklung im Rahmen der medizinischen Dokumentation erklären:

Die zwischenärztliche Kommunikation ist eine der wichtigsten Funktionen der medizinischen Dokumentation. Grundsätzlich unterscheiden wir hier eine patientenbezogene und eine befundbezogene Dokumentation. Erstere ist Realität und fester Bestandteil der ärztlichen Routinearbeit, jede Überweisung erfolgt ja unter Erstellung eines Befundtextes. Die zweite ist die Dokumentation im engeren Sinn. Diese - heute als ein integrierender Faktor jedes wissenschaftlich-medizinisch arbeitenden Institutes zu betrachten - zielt auf die Auswertung von Patientendokumenten irgendwelcher Art (Arztbrief, Krankenblatt, pathologisch-anatomischer Befundbericht und ähnliches) ab.

Es ist bekannt, daß das Auffinden und Auswerten von Befunden aus den Archiven nur unter erheblichem Arbeitsaufwand erfolgt. Derartige befundbezogene Untersuchungen wurden meist nur mit eingeengter Fragestellung - also bei stark reduzierten Testparametern und Untersuchungskollektiven - durchgeführt.

Die Verschlüsselung der nichtnumerischen Information war der erste große Schritt in Richtung Rationalisierung und Ausbau der befundbezogenen Dokumentation. Es entstanden eine Reihe von Schlüsselsystemen, z.B. SNDO, ICD, KDS oder SNOP, um einige zu erwähnen. Natürlich sind durch die festliegende Kapazität der Schlüsselsysteme Grenzen gesetzt, da bei der Schlüsselkonstruktion nicht alle ev. Auswertungsaspekte vorausgesehen werden können. Außerdem muß bei der Datenübertragung ins System - beim Verschlüsseln - nach wie vor bei jedem Probanden eine erhebliche Zusatzleistung zur Routinebearbeitung erbracht werden.

Es sind dies nur zwei Aspekte jener Probleme, die die Dokumentation mittels Verschlüsselung gebracht hat, jedoch erscheint es angesichts der erheblich ausgeweiteten Aufgaben der modernen Medizin immer schwieriger, die Routinekapazität der Kliniken und medizinischen Institute für derartige Vorarbeiten einer befundbezogenen Dokumentation in Anspruch zu nehmen.

Unter diesen Aspekten ist die Idee entstanden, den Computer auch für die Übertragung des Primärtextes in ein Klassifikationssystem und Ordnungssystem zu verwenden. Im medizinischen Bereich findet sich ein solches System erstmals vor etwa 1o Jahren unter dem Namen "automated retrieval" im englichen Sprachbereicht (SMITH and MELTON 1963), später als "natural language retrieval system" (LAMSON 1965) und schließlich auch im deutschen Sprachraum als "vollautomatische Dokumentation" (RÖTTGER 1967 und 1969) sowie "Klartextanalyse" (RÖTTGER 197o, BECKER 1971) englisch übersetzt "free text analysis".

Eine Gegenüberstellung von Verschlüsselung und Klartextauswertungssystem im Diagramm zeigt den Unterschied (Abb.1). Im oberen Teil erkennt man die zweifache ärztliche Leistung (Befunderstellung und Verkodung - denn auch diese muß, soll sie fein genug sein, zumindest unter ärztlicher Kontrolle stehen) vor der Eingabe in die DV-Anlage; das Gitter weist überdies darauf hin, daß bereits hier relevante Information gefiltert wird und für das "storage" verlorengeht. Unten der Routinebefund, der direkt den Datenträger erzeugt. Das " manual coding" wird durch die automatische Standardisierung ersetzt.

Definition der Klartextanalyse

Aus verschiedenen Definitionsversuchen läßt sich heute etwa folgende Zusammenfassung bilden:

Die Klartextanalyse medizinischer Befunde ist ein automatisches Verfahren zum Speichern ("storage"), Bearbeiten ("analysis") und Wiederauswerten ("retrieval") von Textdokumenten des Routinebetriebes (SMITH und MELTON 1963, LAMSON und GLINSKI 1965, RÖTTGER 1967, PRATT 1971 und BECKER 1971).

Diese 3 letzten Teilfunktionen bekommen in der Klartextanalyse überdies einen neuen Gesichtspunkt. Im "retrieval" kann auf die ursprüngliche Form der Diagnose zurückgegriffen werden, da diese ja gespeichert wurde. Zum Unterschied dazu finden sich im Speicher der herkömmlichen Systeme die bereits analysierten, verkodeten Befunde.

Bedingungen und Voraussetzungen der Klartextanalyse

Die Klartextanalyse - und dies soll für alle zum Zeitpunkt funktionsfähigen Systeme gelten - wird unter folgenden Be-

dingungen und Voraussetzungen durchgeführt:

1. Routinebetrieb - die zwischenärztliche, patientenbezogene Dokumentation - bestimmt das Verfahren und nicht umgekehrt, d.h.
   a) das Routinetext-Dokument erfüllt weiterhin eine primäre Funktion, es bleibt "lesbar".
   b) dem datenerhebenden Arzt werden keine terminologischen Auflagen gemacht, bei der Kennzeichnung eines pathologischen Befundes nach Lokalisation, Spezifität und Modifikation kann er sich weiterhin voll auf den zu beschreibenden Sachverhalt konzentrieren.

2. Unter den nichtnumerischen Informationen eines Fallberichtes beschränkt sich das Auswertungssystem auf die Befundauflistung bzw. Diagnosenzusammenstellung. Voraussetzung dafür sind:
   a) Die Befundauflistung als Teil des Routinedokumentes ist maschinenlesbar markiert, d.h. maschinell auffindbar, respektive abgrenzbar.
   b) logisch zusammengehörende Sachverhalte (sog.Einzeldiagnosesätze) sind innerhalb der Befundauflistung voneinander abgrenzbar.
   c) Die Befundauflistung als Aneinanderreihung von Mitteilungen über Sachverhalte nimmt die natürliche Sprache mit ihren Variations- und Kombinationsmöglichkeiten nur begrenzt in Anspruch. Das Auswertungssystem kann auf diese Begrenzung eingerichtet werden.
   d) Das Auswertungssystem wird nicht auf andere, voll in "natürlicher" Sprache abgefaßte Anteile des Fallberichtes (Befunddiskussionen, Epikrisen, Diagnose-Erörterungen an Detailbeschreibungen) ausgeweitet.
   e) Die Kennzeichnung des Verfahrens als "natural language retrieval system" ist nur mit diesem Vorbehalt möglich, da sie sonst zu größeren Hoffnungen auf einer Reihe von medizinischen Teilgebieten Anlaß gibt (LAMSON 1965 und 1967) als später erfüllt werden können (LAMSON 1971, PRATT 1971).

3. Das Auswertungssystem ist auf die komplette Erfassung von Einzeldiagnosesätzen ausgerichtet, in denen auch mehrere logisch zusammengehörende pathologische Befunde enthalten sein können.

   Voraussetzungen dafür sind:

   a) Diagnose- bzw. Befund-Begriffe werden auch dann als Einheit wiedererfaßt, wenn sie im Primärtext durch Aneinanderreihung mehrerer Unterbegriffe definiert waren.
      Hier liegt eine der fundamentalen Voraussetzungen, daß nämlich das System "Adenokarzinom des Magens", "Magenadenokarzinom", "Magenkarzinom mit adenomatösem Charakter" und "adenomatöses Karzinom des Magens" unter dem gleichen Begriff speichern muß.
   b) Homonyme behalten ihre Beziehung zum Kontext bei.

4. Das Auswertungssystem ist variabel. Einzelbegriffe, Einzeldiagnosesätze, Gesamtbefunde je Patient sind sowohl für sich allein als auch in Zuordnung zu Krankheits- und/oder Lokalisations-Überbegriffen ohne Informationsverlust einem auto-

matischen "retrieval" zugänglich.

Voraussetzungen dafür sind:

a) Bezeichnungen im Primärtext werden zu Begriffen im Speichertext überführt. Formal- und Syntax-Varianten sowie Synonyma werden jeweils einem systeminternen Standardbegriff ("perferred term") zugeordnet. Es sollte dies eine in den "storage"-Bereich integrierte Funktion sein.
b) Begriffsimplikationen werden über das Thesaurusregister, das ist der "analysis"-Bereich, erschlossen und als übergeordnete ("superordinate") oder gleichrangige ("correlative") Zusatz-Notationen in die Einzeldiagnosesätze eingefügt.
c) Feinere begriffliche Differenzierungen innerhalb der Diagnosesätze bleiben auch im Speichertext noch erfaßbar.
d) Kein relevantes Wort des Primärtextes wird unterdrückt. Damit können bei der Abfrage durch logische Verneinung negative Aussagen ("kein Anhalt für", "nicht" u.a.) sowie zweifelhafte Aussagen ("verdächtig auf", "Verdacht auf") aussortiert werden.

Diese Voraussetzungen erscheinen im Moment unumgänglich, an ihnen soll ein bestehendes medizinisches Klartextverfahren geprüft werden. Sie sind in manchen Punkten eng gefaßt, in anderen lassen sie einen größeren Spielraum zu, sie decken sich zum Großteil mit den Anforderungen der übrigen Literatur (WONG und GAYNON 1971). Zu betonen ist jedoch, daß praktisch alle Systeme noch im Stadium der Entwicklung und Analyse stehen.

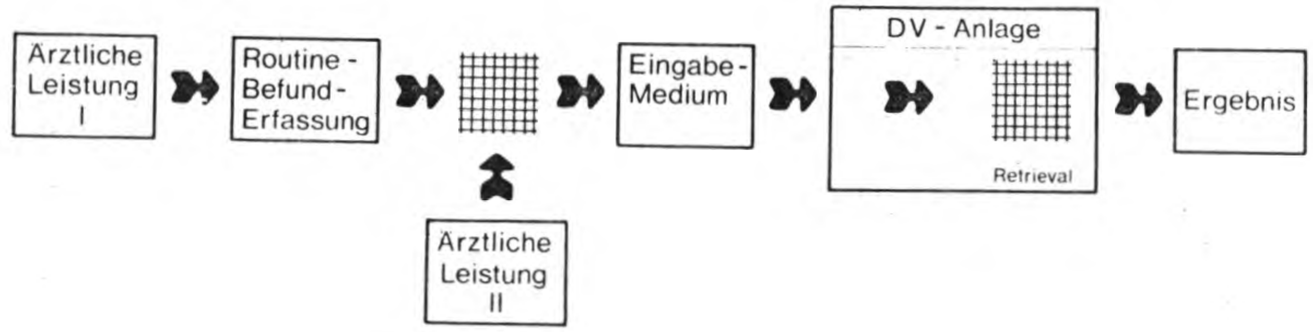

Abb.1: Gegenüberstellung des Aufwandes (ärztliche Leistung) bei Verschlüsselung (oberes Diagramm) und Klartexanalyse (unteres Diagramm).
Ärztliche Leistung I = Erstellung des Befundes
Ärztliche Leistung II = Verschlüsselung

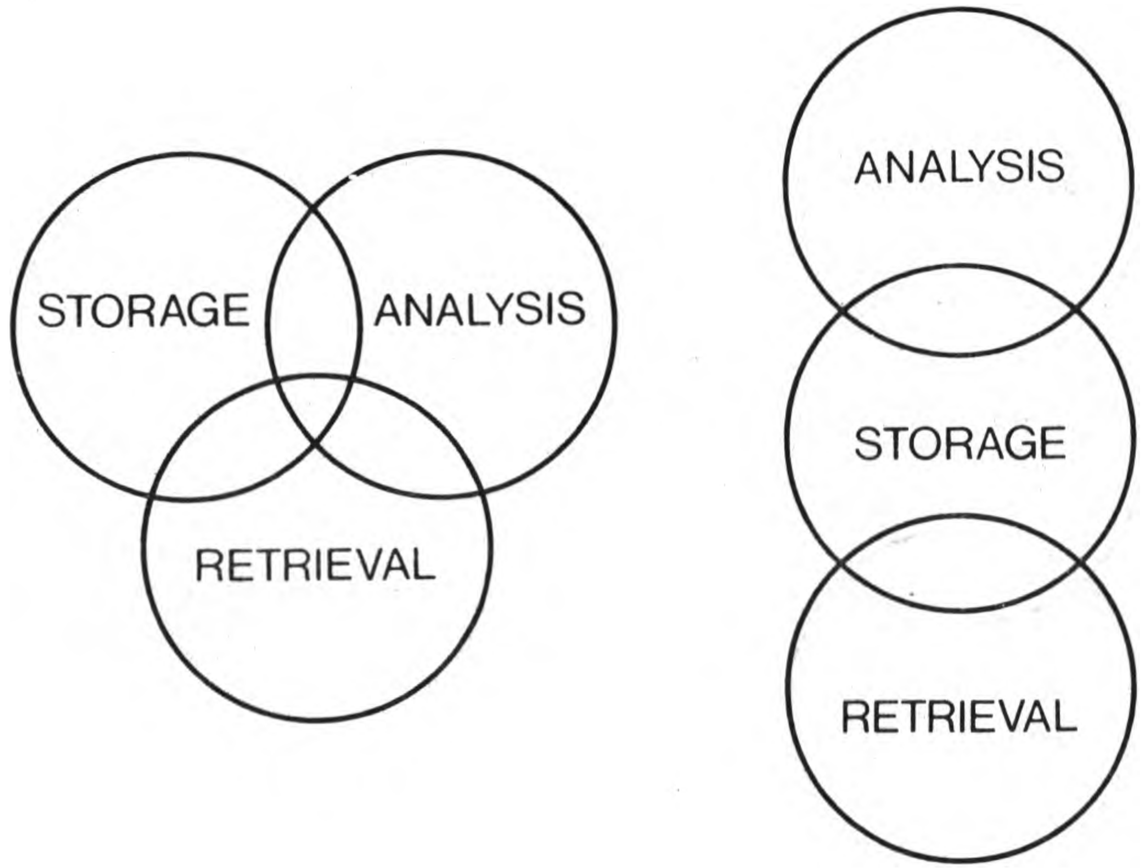

Abb.2: Das zur wissenschaftlichen Auswertung gespeicherte Material (storage) entspricht bei der Klartextanalyse (A) dem ursprünglichen Text und ist sowohl über eine beliebige Analyse als auch direkt der Auswertung zugängig. Bei der Verschlüsselung (B) wird erst das verschlüsselte Material (die jeweilig angewendete Verschlüsselung ist in dem Falle die Analyse) gespeichert.

LITERATURVERZEICHNIS

1) AMERICAN MEDICAL ASSOCIATION (Ed.): Standard Nomenclature of Diseases and Operations. 4.-5. Edition, Chicago 1964/65.

2) BECKER, H., GELL G., SCHWARZ, F., ENGE, H., MUHRI, W.: Klartext-Analyse mit internationaler Klassifikation: Überregionales Pathologie-Register für 29 Krankenhäuser. In G. Fuchs und G. Wagner (Hrsg): Krankenhaus-Informationssysteme. Bericht über die 16. Jahrestagung der Deutschen Gesellschaft für medizinische Dokumentation und Statistik in der DGD vom 3. bis 6. Oktober 1971 in Berlin, S. 247-259 (Schattauer, Stuttgart-New York 1972).

3) COMMITTEE ON NOMENCLATURE AND CLASSIFICATION OF DISEASES (Ed.): Systematized Nomenclature of Pathology (SNOP). (Chicago 1965).

4) IMMICH, H.: Klinischer Diagnosenschlüssel (KDS). (F.K. Schattauer, Stuttgart 1966).

5) LAMSON, B.G.: Storage and Retrieval of Uncoded Tissue Pathology Diagnoses in the Original English Free-Text Form. (Proceedings of the 7th IBM Medical Symposium, Poughkeepsie 1965).

6) LAMSON, B.G. and DIMSDALE, B.: A Natural Language Information Retrieval System. Proceed. IEEE 54: 1636-164o, 1966.

7) LAMSON, B.G., RUSSELL, W.S., FULLMORE, J., NIX, W.E.: The First Decade of Effort: Progress Toward a Hospital Information System at the UCLA Hospital, Los Angeles, California, Meth. Inform. Med. 9 (197o) 73-8o.

8) PRATT, A.W.: Progress towards a Medical Information System for the Research Environment. In G. Fuchs und G. Wagner (Hrsg.): Krankenhaus-Informationssysteme. Bericht über die 16. Jahrestagung der Deutschen Gesellschaft für medizinische Dokumentation und Statistik in der DGD vom 3. bis 6. Oktober 1971 in Berlin, S. 319-336. (F.K. Schattauer, Stuttgart-New York 1o72).

9) RÖTTGER, P.: Diskussionsbemerkungen zu W. Jacob: Moderne Dokumentationsmethoden im Routinebetrieb eines pathologischen Institutes. In G. Griesser und G. Wagner (Hrsg.): Automatisierung des klinischen Laboratoriums. (F.K. Schattauer, Stuttgart-New York 1968).

1o) RÖTTGER, P., REUL, H., KLEIN, I. und SUNKEL, H.: Die vollautomatische Dokumentation und statistische Auswertung pathologisch-anatomischer Befundberichte. Meth. Inform. Med. 8: 19-26, 1969.

11) RÖTTGER, P., REUL, H., KLEIN, I. und SUNKEL, H.: Neue Auswertungsmöglichkeiten pathologisch-anatomischer Befundberichte. Klartextanalyse durch Elektronenrechner. Meth. Inform. Med. 9: 35-44, 197o.

12) RÖTTGER, P.: Die Klartextanalyse pathologisch-anatomischer Befundberichte durch Elektronenrechner. Verh. Dtsch. Ges. Path. 54, 582-588, 197o.

13) SMITH, J. and MELTON, J.: Automated Retrieval of Autopsy Diagnose by Computer Technique.
Meth. Inf. Med. 2, 3 p 85-9o, 1963

14) WONG, R.L., GAYNON, P.: An automated parsing routine for diagnostic statements of surgical pathology reports.
Meth. Inform. Med. lo (1971) 168-175.

15) WORLD HEALTH ORGANIZATION (Ed.): International Classification of Diseases, Volume I and II (8. Revision, Genf 1967 und 1969).

Thesaurus - begriffliche Problematik und strukturelle Information

W.W. Höpker, K. Kayser, W. Ramisch

Pathologisches Institut der
Universität Heidelberg (Prof.Dr.W. Doerr)

Die Entwicklung systematischer Krankheitsklassifikationen und Terminologien in der Medizin darf als Spiegel des begrifflichen Differenzierungs- und Wissenstandes in diesem Fach angesehen werden. Diese hat in der Gegenwart insofern einen gewissen Abschluß gefunden, als es in angemessener Form und in einem praktikablen Umfang und außerdem innerhalb vertretbarer Zeitabstände nicht gelungen ist, den tatsächlichen terminologischen und klassifikatorischen Bedürfnissen auch nur annäherungsweise gerecht zu werden. Folgerichtig ist die Entwicklung automatisierter Klassifikationsverfahren ins Stocken geraten. Die Problematik des zu verarbeitenden sprachlichen Arsenals (sei es von seiner Lexik oder Syntagmatik her) ist bisher grundsätzlich einer formalen Betrachtung in der Medizin verschlossen geblieben; die Diagnosensprache ist als Phänomen bisher kaum theoretisch erörtert worden. Doch verdichten sich die Anzeichen dahingehend, daß auch für diesen speziellen Bereich möglicher EDV-Anwendungen Strukturregeln ähnlicher Komplexität angenommen werden müssen, wie sie mit der generativen Grammatik bereits für die natürliche Sprache formuliert wurden.

Von einem Gesamtkonzept sind wir weit entfernt. Hier soll nur schlaglichtartig Abgrenzung und Funktionsweise unterschiedlicher Wörterbücher erörtert und eine Terminologie dieser Begriffe vorgeschlagen werden. Spricht heute jemand von einem "Thesaurus", so reichen die Vorstellungen der Zuhörer von einer "großen Liste" bis zu einem "terminologisch geordneten und lexikalisch operationalisierten Wortverzeichnis". - Weiterhin wird gefragt, ob von Seiten der Informationstheorie formale Anhaltspunkte als konkrete Hilfestellung bei der Konstruktion eines Thesaurus erwartet werden können.

In Abb. 1 sind Begriffe wie "Schlüsselliste","Begriffsliste", "Dokumentationssprache","Klassifikationssystem" und "Thesaurus" einander gegenübergestellt. Ausgangspunkt ist eine lose Sammlung von Begriffen, die Begriffsliste. In dieser sollen alle diejenigen Bezeichnungen erscheinen, die in der betreffenden medizinischen Fachsprache benutzt werden und mit einem operationalisierbaren Inhalt (wenn auch nur im Sinne der Diagnose als konkrete Handlungsweisung für den Arzt) belegt werden können. Der erste Bearbeitungsschritt an einer solchen Begriffsliste führt zur sog. Schlüsselliste. Meist wird die Begriffsliste einer Revision nach (angenähert) terminologischen oder pragmatischen Gesichtspunkten unterworfen. Eine Schlüsselliste hat keine anderen Merkmale als die, nach dem Prinzip der "Ablage" und des "Wiederfindens" eine äußerliche (und somit nicht inhaltlich bestimmte) Ordnung zu schaffen. "Schiefrige Induration der Lunge" und "alte, vernarbte Lungenspitzentuberkulose" erscheinen ohne weiteres gleichzeitig nebeneinander. Zugang und Abgang von Begriffen ist meist nicht streng vorgeschrieben.

Eine Schlüsselliste jedoch, in der man "alles wiederfindet", ist Voraussetzung für die weiteren Differenzierungsschritte. Unter den allgemeinen Gesichtspunkten der Dokumentation werden Vorzugsbenennungen (sog. Preferred Terms) definiert. Dies kann unter ausschließlich inhaltlichen Gesichtspunkten (z.B. aktive und inaktive Tuberkulose) geschehen. Andererseits können Beziehungs- und

Verknüpfungsvorschriften (Relatoren) angegeben werden. Aber auch die Aufnahme von Konstruktionsregeln (welcher Begriff unter welchen Bedingungen aus der Schlüsselliste übernommen werden soll) sind als formale Anweisungen zur Abgrenzung der Vorzugsbenennungen anzusehen.

Sachlogische Beziehungen hingegen sind die Voraussetzungen von Klassifikationssystemen, welche generativ oder assoziativ strukturiert sein können. Diese erzeugen thematische Gruppen mit jeweils mehrdimensionaler oder hierarchischer Zuordnung.

Ein optimaler Zugang zu einem Klassifikationssystem ist eigentlich erst dann möglich, wenn zu diesem gleichzeitig Schlüsselliste und Begriffsliste mit angegeben werden. In den Fällen, in denen diese explizit nicht mitaufgeführt werden, werden sie stillschweigend vorausgesetzt. Die operationalisierte und damit gebrauchfähige Erweiterung eines Klassifikationssystemes ist demnach ein Verzeichnis, in welchem zu den Vorzugsbenennungen auch die Nichtvorzugsbenennungen aufgelistet sind. Die formale "Instanz", welche zwischen Vorzugsbenennungen und Nichtvorzugsbenennungen unterscheidet, wird allgemein als "terminologische Kontrolle" bezeichnet (wobei allerdings nicht die Terminologie im konventionellen Sinne angesprochen wird). Die Systematik und die sachlogischen Beziehungen werden aus dem entsprechenden Klassifikationssystem entnommen. Erst dann, wenn sämtliche vier Kriterien gegeben sind, sprechen wir von einem Thesaurus im engeren Sinne.

An einen medizinischen Thesaurus (besser: Thesaurus in der Medizin) werden noch weitergehende Anforderungen gestellt:

1. Dieser muß - zumindest in gewissen Bereichen - zu anderen medizinischen Disziplinen und Fachbereichen kompatibel sein. Beispiel: Hypertonie und Diabetes sind Allgemeinerkrankungen, die auch für den Ophthalmologen von Bedeutung sein können; in einem "rein ophthalmologischen Thesaurus" erscheinen diese Begriffe nicht. Wie sonst sollen sie abgebildet werden, als durch entsprechenden Zugang zur jeweiligen Nachbardisziplin?
2. Zusätzlich soll Kompatibilität zu den großen internationalen Schlüsselsystemen wie ICD und SNOP gegeben sein.
3. Ein medizinischer Thesaurus soll nicht nur als "Codierungsreservoir" mit den entsprechenden Anweisungen und Regeln verstanden werden. Vielmehr ist er als Knotenpunkt des gesamten Informationsflusses in der jeweils zu versorgenden Organisationseinheit anzusehen. Neben der Codierung soll die Funktion als Schlagwortkartei (Bibliothek, Sammlung) und Nachschlagwerk (in der fachlichen Aus- und Weiterbildung) zusätzlich wahrgenommen werden.

Ohne Zweifel stellt auch ein Thesaurus nur ein nach bestimmten Regeln strukturiertes Verzeichnis dar - stellt man den Vergleich mit der eingangs erwähnten Diagnosensprache an. Zur Diagnosensprache gehören nicht nur sachlogische Beziehungen (wie sie in einem Thesaurus ihren Niederschlag finden) sondern in Erweiterung der terminologischen Kontrolle auch die formalen und semantischen Beziehungen der Sprachlogik (im Sinne der Syntagmatik resp. Grammatik). In dieses Feld gehört auch die Gesamtzahl der kontextlichen Beziehungen. Unter Kontext verstehen wir Sprachstrukturen, die nichtinhaltlich innerhalb eines oder mehrerer Wörter definiert werden können, jedoch rückwirkend den Bedeutungsinhalt dieser Wörter beeinflussen und

determinieren können. Es ist lange überhaupt bezweifelt worden, daß derartige Beziehungen auch für die Diagnosensprache von Bedeutung sein könnten. (Hiervon kann man sich jedoch schnell überzeugen, wenn man gelegentlich ein Krankenblatt oder ein Sektionsprotokoll z.B. aus dem Jahre 195o zur Hand nimmt. Auch heute noch sind sie uns durchaus "verständlich", doch können wir im einzelnen oftmals nicht angeben, welche konkrete Vorstellung bei der Niederschrift dahinter gestanden haben mag.)

In der zweiten bereits angesprochenen Frage soll erörtert werden, ob wir aus anderen Wissenschaftbereichen wie z.B. der Informationstheorie Hilfestellung bei der Konstruktion eines Thesaurus erwarten können. - Das bereits seit etwa 4o Jahren in die Linguistik eingeführte Rangzahlprinzip hat auch Eingang in die Informationstheorie gefunden. Nach diesem können nicht nur Benutzungshäufigkeiten (entsprechend der ursprünglichen Definition) sondern auch Gliederungen und Klassifikationen nach ihrer Größe miteinander verglichen werden.

In Abb. 2 haben wir nach diesem sehr einfachen Prinzip die Benutzungshäufigkeit von Wörtern der deutschen Schriftsprache mit denjenigen zweier Dokumentationssprachen verglichen. Auffallend ist, daß der Verlauf der Rangzahlkurve beider Dokumentationssprachen "flacher" und vor allem "gekrümmt" im Vergleich zur natürlichen Sprache erscheint. Offensichtlich bestehen doch deutliche Unterschiede (zumindest im Hinblick auf die jeweilige Rangzahlverteilung) zwischen der natürlichen und der beiden Dokumentationssprachen. Lassen sich nun weitere informationstheoretisch ableitbare Merkmale finden, die diese Beobachtung erhärten können?

Betrachtet man die Wörter der ersten Rangzahlpositionen der natürlichen Sprache nach ihrer inhaltlichen Bedeutung, so stellen wir fest, daß diese einen relativ geringen Informationsgehalt besitzen. Es sind dies in der Umgangsprache die Wörter "und", "auf" ec., in der Dokumentationssprache "schlaffe Dilatation des Herzens", "Hyperämie der parenchymatösen Organe" ec. Der partielle Informationsgehalt einer Nachricht ist dem Logarithmus dualis seiner relativen Häufigkeit umgekehrt proportional. Somit nimmt der Informationsgehalt einer Nachricht mit seiner Häufigkeit ab. Als Entropie bezeichnen wir den Mittelwert sämtlicher partieller Informationsgehalte. Da die Differenz aus Entropie und relativer Entropie als Redundanz definiert ist (beide sind demnach mittelwertsabhängig) ist folgende Situation gegeben:

1. Trotz der unterschiedlichen Rangzahlkurven kann die Entropie für beide - die natürliche wie die Dokumentationssprache - gleich sein (wie man dem gekreuzten Verlauf beider Kurven leicht entnehmen kann);
2. auch kann der gesamte Informationsgehalt gleich sein, obwohl zwei völlig verschiedene strukturierte Systeme zugrunde liegen.

Ohne Zweifel stoßen wir hier an die Grenzen der heute gebräuchlichen und wesentlich von SHANNON inaugurierten Informationstheorie. Mittelwertsabhängige Größen als alleinige Informationsmaße versagen bei differenzierteren Fragestellungen.

An einem weiteren Beispiel soll die Relativität dieses Mittelwertmaßes als Informationsgröße demonstriert werden (Abb. 3). Die ver-

schiedenen Schlüsselsysteme (ECD (3-stellig); PADS (Becker); ICD (4-stellig); KDS (Immich); PDS (Heidelberg)) wurden entsprechend der Größe ihrer jeweiligen Klassenaufteilung einander nach Rang gegenübergestellt. Während ICD und PADS noch eine angenäherte Gleichverteilung aufweisen, haben sich KDS und PDS immer weiter von einer solchen idealen Verteilung entfernt. Dieser sehr eindruckvolle Unterschied kann durch Gegenüberstellung der relativen und maximalen Informationsentropie anschaulich gemacht werden derart, daß mit zunehmender "maximal" darstellbarer Information die auf das Inventar bezogene und damit tatsächlich benutzte Information bei den betrachteten fünf Schlüsselsystemen abnimmt (Abb. 4).

Somit müssen wir die Frage nach einer Hilfestellung von Seiten der Informationstheorie bei der Konstruktion eines Thesaurus abschlägig beantworten.

Da weder die eingangs gegebenen Definitionen noch die oben kurz skizzierten Informationsmaße uns in die Lage versetzen, einen Thesaurus optimal zu konstruieren, müssen wir fragen, welche Argumente überhaupt für oder gegen eine bestimmte Thesaurusgliederung sprechen können. Hier sind unserer Ansicht nach ausschließlich pragmatische Gesichtspunkte von Bedeutung wie sie bereits für den medizinischen Thesaurus genannt wurden: Kompatibilität zu anderen Fachbereichen und bereits bestehenden und gebräuchlichen Schlüsselsystemen, umfassende und breite Anwendungsmöglichkeiten in sämtlichen relevanten Informationsbereichen.

Unsere Standortbestimmung könnte folgendermaßen lauten:

1. Begriffsliste, Schlüsselliste, Dokumentationssprache, Klassifikationssystem und Thesaurus sollten einheitlich benutzt werden;
2. Voraussetzung eines klinischen Informationssystemes ist der Thesaurus. Dieser soll inhaltlich und formal dem zu versorgenden Informationsbereich voll genügen;
3. Selbstverständlich soll ein Thesaurus geringe "Kosten" (im informationstheoretischen und finanziellen Bereich) verursachen;
4. die bekannten informationstheoretischen Kriterien reichen nicht aus, wesentliche neue Impulse bei der Thesaurusgestaltung zu geben;
5. um so mehr muß - um Fehlentwicklungen größeren Ausmaßes zu vermeiden - auf die vielfältigen Anwendungsbereiche mit den jeweiligen pragmatischen Aspekten verwiesen werden.

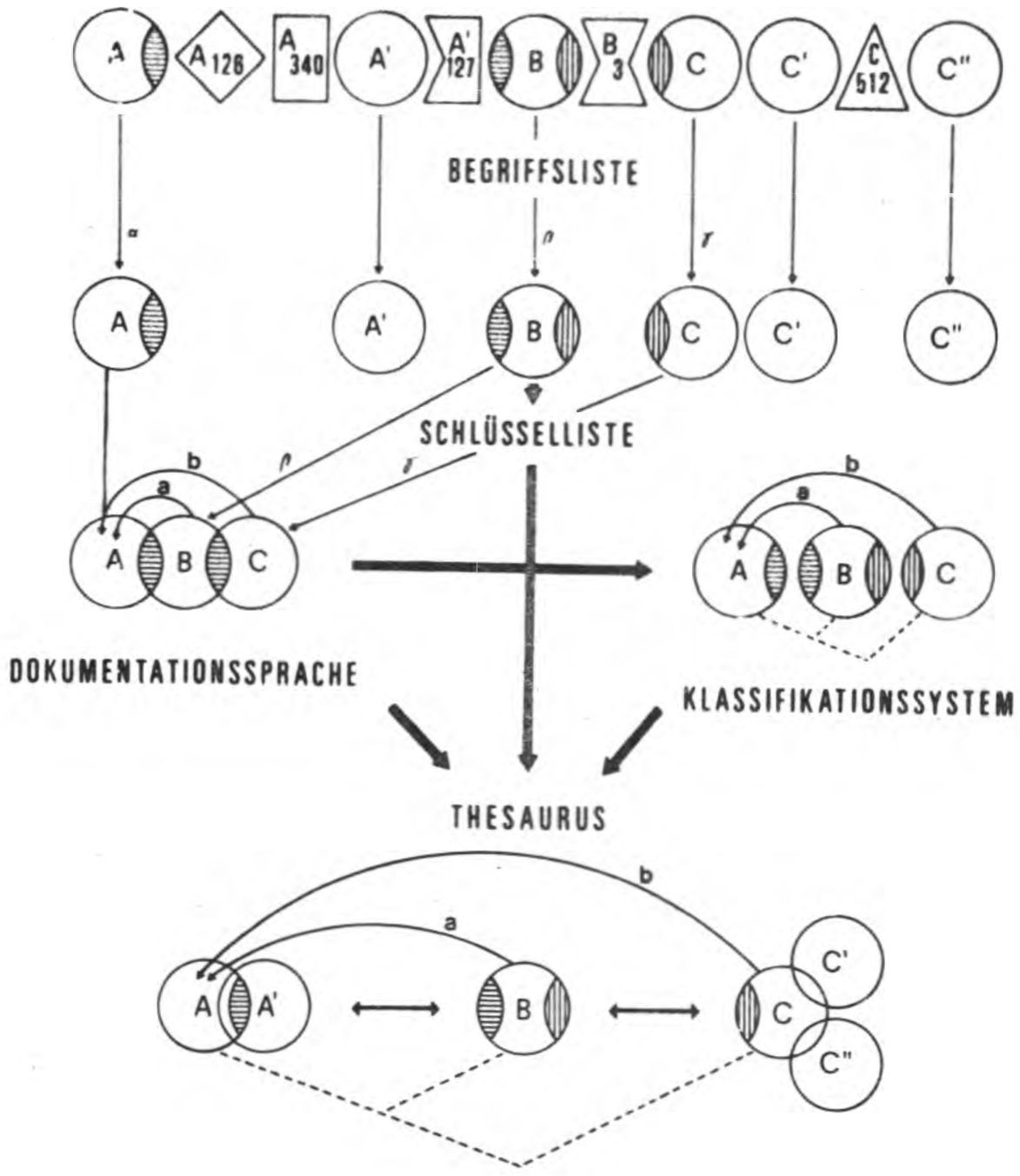

Abb. 1

Gegenüberstellung von "Begriffsliste", "Schlüsselliste", "Dokumentationssprache", "Klassifikationssystem" und "Thesaurus". Die Begriffsliste besteht aus einer Auslistung (rohen Sammlung) von Begriffen eines Fachbereiches. Der Übergang von den Begriffen zu den Begriffsbenennungen eliminiert einen großen Teil der ursprünglich gesammelten Begriffe, fügt aber auch neue hinzu (nicht dargestellt, Schlüsselliste). Zur Dokumentationssprache gehören Konstruktionsregeln (,ß,) und eine Liste von Relatoren (a,b). A,B und C sind Vorzugsbennungen, (Quasi-)Homonyme und (Quasi-)Synonyme sind bekannt und angegeben. Zu einem Klassifikationssystem tritt die Systematik der sachlogischen Beziehungen hinzu (durch die hierarchische Verzweigung gekennzeichnet). Aus jedem der drei Zwischenstufen (Dokumentationssprache, Schlüsselliste, Klassifikationssystem) kann ein Thesaurus (mit den Nichtvorzugsbennungen A', C' und C'') erstellt werden.

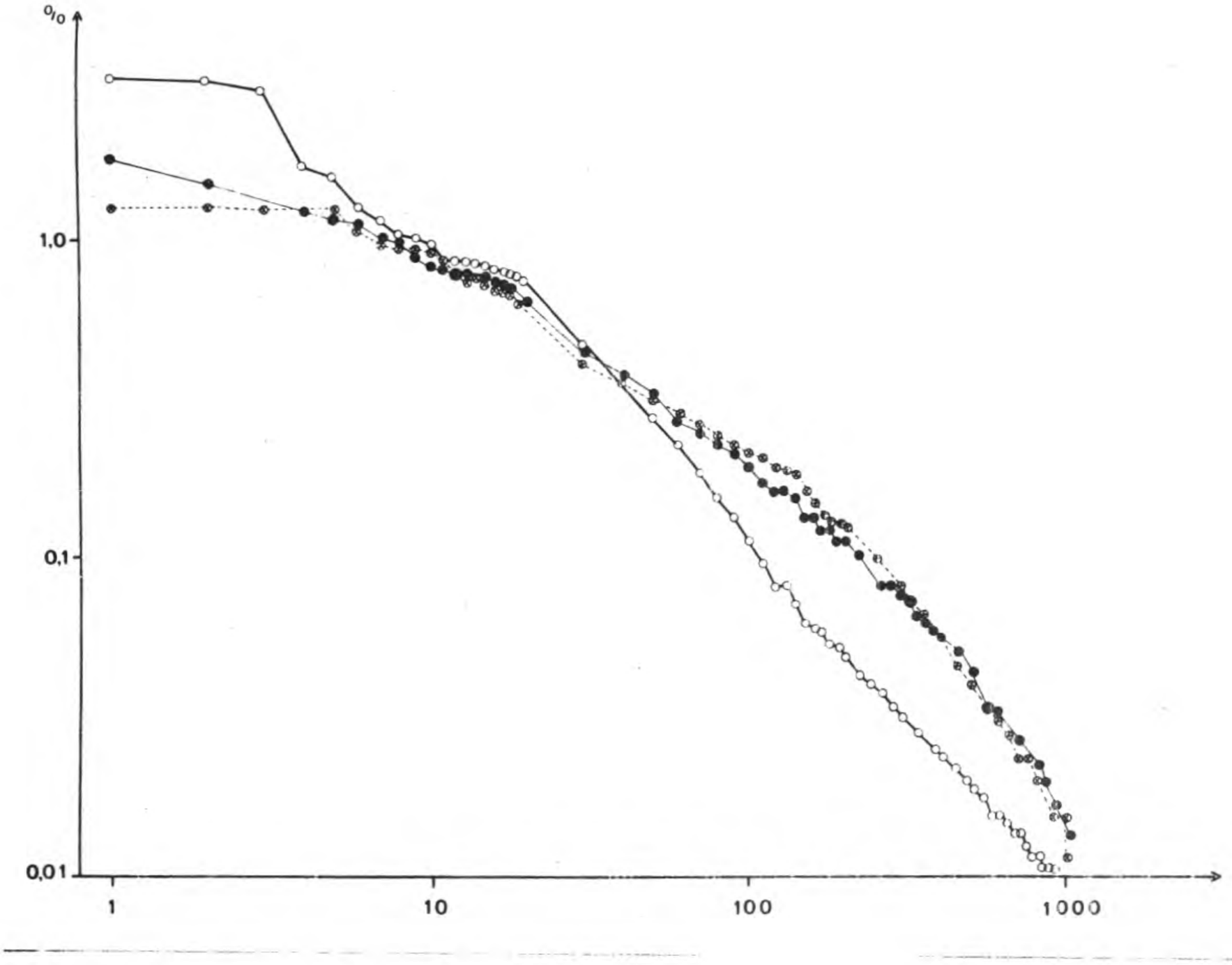

Abb. 2

Rangzahlkurven relativer (Benutzungs-)Häufigkeiten. Ausgezogene Kurve mit leeren Kreisen nach Angaben von HELMUT MEIER. Es handelt sich hier um die KAEDING'schen Zählungen von $1.09 \cdot 10^6$ Wörtern der deutschen Prosa um das Jahr 1900 (neuere Zählungen stehen uns nicht zur Verfügung). Ausgezogene Kurve mit gefüllten Kreisen: Diagnosensprache des Pathologischen Institutes A. Unterbrochene Kurve mit "x" gefüllten Kreisen: Diagnosensprache des Pathologischen Institutes B. Ordinate: relative Häufigkeit in Prozent. Abszisse: Entsprechende Rangpositionen in logarithmischem Maßstab.

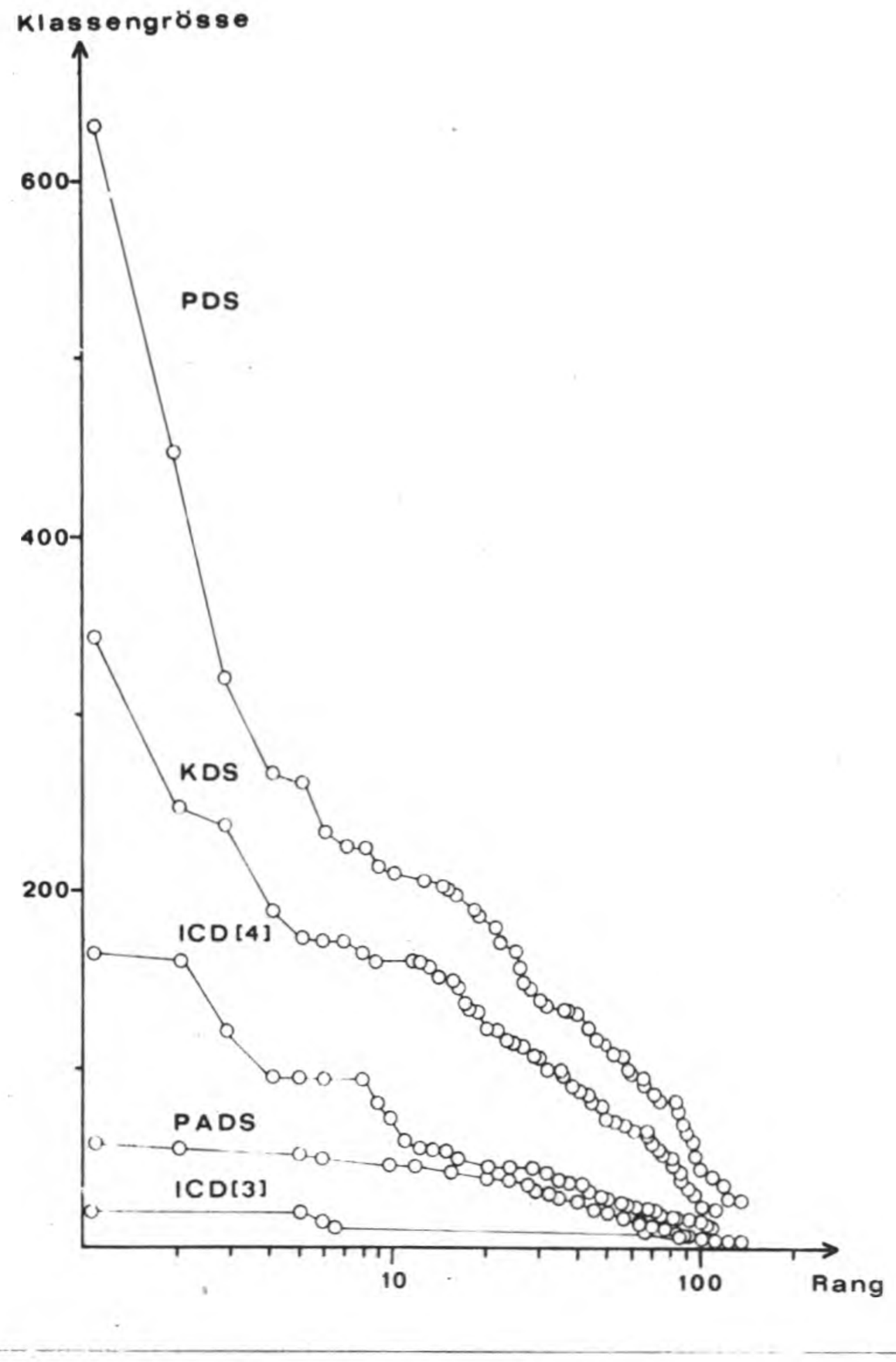

Abb. 3

Rangzahlkurven absoluter Klassengrößen von fünf verschiedenen Schlüsselsystemen. Ordinate: Klassengröße; Abzisse: Rang in logarithmischem Maßstab. ICD (3): 3-stellige International Clasdification of Diseases; PADS: Pathologisch-anatomischer Diagnosenschlüssel (Becker, Graz); ICD (4): 4-stellige (jetzte gebräuchliche) ICD; KDS: Klinischer Diagnosenschlüssel von Immich; PDS: Pathoanatomischer Diagnosenschlüssel (Heidelberg). Die mehrdimensionalen Schlüsselsysteme wie z.B. SNOP und SNDO sind mit den präkoordinierten Systemen nicht vergleichbar und konnten deshalb nicht mit aufgenommen werden.

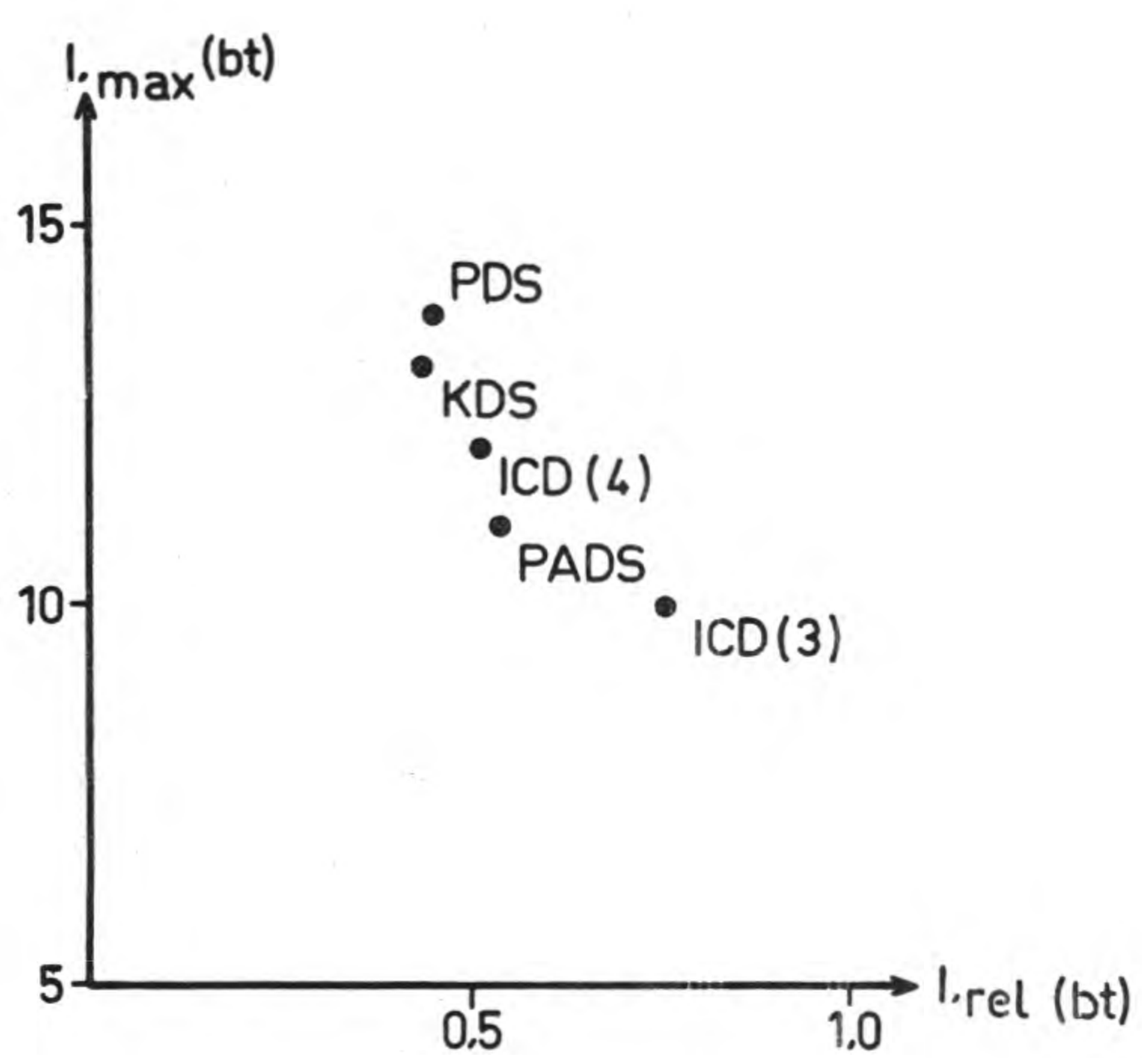

Abb. 4

Maximale Informationsentropie (I,max; Ordinate) und relative Entropie (I, rel; Abzisse) bei den fünf Schlüsselsystemen. Die relative Entropie entspricht dem tatsächlich genutzten, die maximale Entropie dem von der Aufteilung her möglichen Informationsraum. Mit steigender maximaler Entropie sinkt die relative Entropie.

Klartextsysteme in der Medizin außerhalb des deutschen Sprachbereiches

W. Feigl

Pathologisch-Anatomisches Institut der
Universität Wien

Den beiden zum Zeitpunkt (Juni 1973) im deutschen Sprachraum bekannten Klartextanalysesystemen (RÖTTGER 1967 und BECKER u.a. 1971) stehen sechs teils ältere Systeme im außerdeutschen Sprachgebiet gegenüber. Fünf davon wurden für die englische Sprache, eines für das Französische entwickelt.
Von der Anlage her lassen sich diese in sogenannte Wort-Wort und Wort-Schlüsselsysteme unterscheiden. Erstere speichern den Text Wort für Wort als Code unabhängig von weiteren Kontextzusammenhängen, letztere transferieren die eingehenden Worte gleichzeitig in ein präformiertes Schlüsselsystem (mit den entsprechenden Implikationen und Beschränkungen). Unter den englischen Systemen gehören drei der ersten Gruppe an.

Das System von LAMSON und DIMSDALE ist am höchsten entwickelt, jenes von WONG und GAYNON besitzt eben noch einen Thesaurus, während jenes von PAPLANUS wie auch das von WOLFF-TERROINE (das erwähnte einzige französische System) zur Abfrage nur mehr ein Wörterbuch besitzen.
Das System von SMITH und MELTON ist ein formales Wort-Schlüsselsystem, de facto jedoch ein Wort-Wortsystem. Das Verfahren von PRATT gehört der zweiten Gruppe an und verwendet den SNOP, wobei es zugleich als einziges Ansätze zur Mitverwendung der Semantik zeigt.

1. Das "Automated Retrieval" von SMITH und MELTON.

   Dieses System wurde erstmals 1963 beschrieben, in der Folge noch in den Jahren 1964 und 1966. Seit damals existiert keine neue Publikation mehr darüber.
   Es wurde dort nach der Durchführung einer empirischen Textanalyse aus Sektionsberichten und Begriffen des SNDO ein Wortcode geschaffen, der ein Speichern und Aufsuchen im Computer ermöglicht. Dieser Code teilt sich in "pathological process" "site" und "modifier", einer der wesentlichesten Gedanken aller folgenden Systeme, wobei die "modifier" wieder Unterhierarchien für diverse Kategorien besitzen. Die Eingabe erfolgt über Lochkarten. Über "retrieval"-Ergebnisse und über den weiteren Ausbau des Systems ist nichts bekannt geworden; seine Bedeutung liegt in der erstmaligen Erörterung einiger Grundelemente medizinischer Befund-Klartext-Analyse, eben vor allem die Unterteilung von NOSOLOGIE, LAGE und MODIFIZIERENDEN BEGRIFF in Bezug auf automatische Auswertung.

2. Das "Natural Language Retrieval" - System von LAMSON, GLINSKI, DIMSDALE.

   1965 und 1966 beschrieben, 197o erschien ein gedruckter Thesaurus. Dieses System wertet chirurgische Biopsie-Dokumente aus. Die Worte werden hier nicht nur in ihrer Primärform sondern zusammen mit automatisch eingebrachter Zusatzinformation gespeichert. Hierzu wurde ein Thesaurus mit einer Horizontalstruktur ("synonym classes") und einer Vertikalstruktur ("subordination", "linkage" und wenigen Querver-

bindungen) erstellt.
Dieser Thesaurus, der keine - und das ist hier wesentlich - durchgehende Hierarchie besitzt, umfaßte nach der letzten Arbeit 197o 1o.ooo Worte, davon etwa 6.ooo Synonymverknüpfungen und 5.ooo andere - meist "subordination"-Verknüpfungen. Die Auswahl dieser "linkages" erfolgt offensichtlich nach Bedarf und kann jederzeit ergänzt werden.
Als Vorteil erscheint hier, daß jedes Wort auch im Speichertext verbleibt und als Wort wieder auffindbar ist. Dadurch wird ermöglicht, daß der befunderhebende Arzt auch die Abfragung konzipieren kann.
Dieses Verfahren ist sehr praktisch eingerichtet. Es dient einem raschen Zugriff zu einem umfangreichen, bezüglich der Textvarianten aber relativ einheitlichen und überschaubaren Datenbestand. Seine Praxisorientierung zeigt sich auch in der a priori durchgeführten Kombination der Datei mit einem Abrechnungsregister. Das "retrieval" scheint sich vorwiegend auf das Wiederauffinden von Einzelfällen und kleinere Fallgruppen zu erstrecken. Über umfangreiche, mit Statistiken kombinierte systematische Abfragungen ist nichts berichtet worden.

3. Das System für "Autopsy Diagnosis Storage und Retrieval without Numerical Coding" von PAPLANUS et al.

Dieses System wurde 1969 erstmals beschrieben. Die Autopsie-Befundauswertung erfolgt an einem gesondert erstellten Zweitdokument des Routinebefundes. Eine Art Punkt-Punkt-Regel zur Abgrenzung zusammengehörender Sachverhalte wird in diesem Zweitdokument eingeführt, wobei ein "slash" (Gedankenstrich") an Stelle eines Punktes tritt. Bestimmte Worte werden als "descriptors" ausgewiesen.
Mit diesen und der Hilfe von Suchprogrammen werden innerhalb des "retrieval" fortlaufend Listen erstellt. Die Aufsuchung erfolgt mit dem Erfassen der Diagnose-Nummer. Bezüglich der Apparatausstattung ist das System für mehrere Benutzer vorgesehen.
Es wird geltend gemacht, daß das System dem Spezialbedarf eines pathologischen Institutes zu genügen hätte, daß jede Beschränkung auf einen fest definierten Wortschatz entfalle und das im Speichertext auch die Wortnuancen erhalten blieben. Der Darstellung nach ist auch dieses Wort-Eingabe-Wort-Ausgabe-System auf die Deckung eines dringend notwendigen Routinebedarfes hin konzipiert worden. Eine begrenzte Dimension des "information retrieval" ist dabei in Kauf genommen worden.

4. Das "Automatic Encoding System" von PRATT.

Von PRATT un PACAK wurde 1968 erstmals auf die Möglichkeit einer semantischen Analyse medizinischer Texte aufmerksam gemacht. 1971 wurde das System von PRATT erstmals beschrieben. Es ist sicherlich eines der interessantesten und wohl auch das aktuellste. Zwei wesentliche Aspekte zeichnen es aus.

1) Wird als Grundlage der SNOP verwendet. Umfangreiche Thesauri werden aber vom Autor abgelehnt, lediglich die wichtigsten pathologisch-anatomischen Begriffe finden sich als Wortstämme samt SNOP-Nummer im Register. Damit

werden 85 % der Befunde erfaßt.

2) Wird durch ein umfangreiches Programm zur Erkennung der Wortstämme, das z.B. "laryngeal" in Larynx umwandelt, das oben erwähnte Wörterbuch möglichst klein gehalten. Die vollständige algorhythmische Aufzweigung von Endungen umfaßt z.B. 95 Varianten. Damit wird nun bis 95 % richtiges "retrieval" erreicht.
Dies soll an dieser Stelle nur ein Überblick sein, eine genauere Darstellung erfolgt von Mitarbeitern der AGK, die sich momentan zum Studium des Systems in Amerika aufhalten.

5. Das französische System.

Bereits 1969 erschien von WOLFF-TERROINE ein medizinischer Thesaurus der gut durchkonzipiert ist, jedoch offensichtlich kaum zur Anwendung kam. Daran schließen weitere Arbeiten der jüngeren Zeit an, die die sprachliche "Umgebung", den Kontext der diversen Begriffe beleuchten.
Wesentlich weniger aufwendig erscheint die Anwendung in der Verarbeitung histopathologischer Daten des GUSTAVE-ROUSSY-Krebsforschungsinstitutes. Aus der histologischen Diagnose werden vom befundenen Arzt nach Art der Beschlagwortung wesentliche Begriffe unterstrichen. Diese werden zusammen mit den anderen Daten abgelocht.
Nach der Eingabe von etwa 1o.ooo solcher Diagnosen wurde ein Wörterbuch erstellt, das die 43o häufigsten pathologisch-anatomischen Bezeichnungen aus dem Gebiet der Cancerologie enthält. Der Begriff "Thesaurus" wurde dafür nicht gebraucht. Lokalisationen sind nicht enthalten und werden vom System offensichtlich nicht berücksichtigt. An Hand dieses Systems können einfache Abfragen gemacht werden, "retrieval"-Ergebnisse wurden bis jetzt nicht gebracht.
Das System stellt offenbar ein Minimalwörterbuch für nosologische Begriffe der Cancerlogie dar, es gilt (was auch betont wird) nur für die Terminologie eines eng umgrenzten Raumes.
Interessant ist der Zuverlässigkeitsindex, der jedem Befund beigegeben wird (sicher, wahrscheinlich, unsicher) und damit ein wesentliches Problem der Klartextanalyse hintanhält.

6. Die "Automated Natural Language Parsing Routine" von WONG und GAYNON.

Dieses System wurde 1972 erstmals publiziert. Die Dateneingabe erfolgt in Gestalt unveränderter Dokumente des Routinebetriebes. Das System befaßt sich mit der Dokumentation der Biopsieberichte ("surgical pathology"). Es ist einmal ausgerichtet auf einfachen Befundzugriff (Sammlungsaspekt), für spätere Phasen sind auch korrelierende Abfragungen mit klinischen Untersuchungen sowie statistische Auswertungen vorgesehen. Der Gebrauchswert des Systems ist vor allem auf eine rasche on-line-Information ausgereichtet. Die Texte in den Dokumenten sind relativ kurz, die auswertbaren Dokumentteile schließen klinische Informationen mit ein.

In der Praxis werden die Formulare auf Magnetbandschreiber geschrieben.
Die Texte werden über ein Wörterbuch ("Lexikon") verarbeitet, dieses selbst ist kontextfrei, "pointer" verbinden die Wortbegriffe mit ihren Implikationen. Bei diesen wird zwischen "word standardisation" und "word extension" unterschieden. Erstere reduziert die Masse des "input" auf "retrieval"-bezogene "output"-Begriffe (Standardisierung) letztere fügt Begriffsimplikationen ein (z.B. Niere bei Glomerulonephritis). Eine weitere feinere Thesaurusdurchstrukturierung wird nicht beschrieben.
Interessant erscheint, daß die Autoren in diesem Zusammenhang eine semantische Analyse der Diagnosen ablehnen, wiewohl sie in einer früheren Arbeit (1 Jahr zuvor) dies in höchst eindrucksvoller Weise durchführen. Dabei werden durch eine Anzahl von "delimiters" (insgesamt 62) 83 % der Diagnosen richtig in Lokalisation, Diagnose und "modifier" unterteilt. Dies gewinnt an Bedeutung, da die Autoren selbst dann das linquistische Problem der "cross over error" herausstellen, wie z.B. im Falle des Diagnosesatzes "Adenokarzinom der "Vagina" und "Rektum" trennen. Die Kontrolle auf bisher unbekannte Worte erfolgt über einen "scanner". Das Wörterbuch enthält 9.ooo Einheiten. Vom laufend dokumentierten Material wird sowohl ein Namens- als auch ein Befundregister erstellt ("inverted file").
"Restrieval Tests" sind an Speicherungen von 3-Monatstextmengen durchgeführt worden. Als Ergebnis wird hervorgehoben, daß das System auch den Ansprüchen genüge, von denen man bisher glaubte, sie seien nur durch Verschlüsselung zu befriedigen.

Abschließend sollte gesagt werden, daß diese momentan bestehenden Systeme einige Kriterien der medizinischen Klartextanalyse nur zum Teil erfüllen. Teils sind es reine Wortabfragesysteme, teils besitzen sie einen strukturierten Thesaurus. Interessant ist, daß echte größere Auswertungen bis heute noch kaum publiziert sind.

LITERATUR

1) BECKER, H., GELL G., SCHWARZ, F., ENGE, H., MUHRI, W.:
Klartext-Analyse mit internationaler Klassifikation: Überregionales Pathologie-Register für 29 Krankenhäuser. In G.Fuchs und G. Wagner (Hrsg): Krankenhaus-Informationssysteme. Bericht über die 16. Jahrestagung der Deutschen Gesellschaft für medizinische Dokumentation und Statistik in der DGD vom 3. bis 6. Oktober 1971 in Berlin, S.247-259 (Schattauer, Stuttgart-New York 1972).

2) DIMSDALE, B.: User's Manual I - A Natural Language Information Retrieval System.
(Los Angeles Scientific Center 35, o19, 1966).

3) GAYNON, P., WONG, R.L.: A Retrieval System for a Library of Pathology Reports, Slides and Kodachromes.
Meth. Inf. Med. 11 (1972) 152-163.

4) GERARD-MARCHANT, R., WOLFF-TERROINE, M., RENAUD, M. and ACARO, M.:
Processing of Histopathological Data.
Meth. Inform. Med. 1o (1971) 142-147.

5) LAMSON, B.G.: Storage and Retrieval of Uncoded Tissue Pathology Diagnoses in the Original English Free-Text Form.
(Proceedings of the 7th IBM Medical Symposium, Poughkeepsie 1965).

6) LAMSON, B.G. and DIMSDALE, B.: A Natural Language Information Retrieval System. Proceed. IEEE 54: 1636-164o, 1966.

7) LAMSON, B. and DIMSDALE, B.: Pathology Thesaurus.
UCLA Hospital - IBM 197o.

8) PAPLANUS, S. et al.: A Computer-Based System for Autopsy Diagnoses Storage and Retrieval Without Numerical Coding.
(Lab. Invest. 2o (1969) 135-14o.

9) PRATT, A.W.: Progrss towards a Medical Information System for the Research Environment. In G. Fuchs und G. Wagner (Hrsg.):: Krankenhaus-Informationssysteme. Bericht über die 16. Jahrestagung der Deutschen Gesellschaft für medizinisch Dokumentation und Statistik in der DGD vom 3. bis 6. Oktober 1971 in Berlin, S. 319-336. (Schattauer, Stuttgart New York 1973).

1o) PRATT, A.W., PACAK, M.: Indentification and Transformation of Terminal Morphemes in Medical English.
Meth. Inform. Med. 8 (1969) 84-9o.

11) PRATT, A.W. and L.B. THOMAS: An Information Processing System Pathology Data.
Pathology Annual-66. (Appleton-Century-Crofts Publisher 1967)

12) RÖTTGER, P.: Diskussionsbemerkungen zu W. Jacob: Moderne Dokumentationsmethoden im Routinebetrieb eines pathologischen Institutes.
In G. Griesser und G. Wagner (Hrsg.): Automatisierung des klinischen Laboratoriums (F.K. Schattauer, Stuttgart-New York 1968).

13) SMITH, J.C.: Anatomic pathology and data processing.
Arch. Path. 81 (1966) 279-28o.

14) SMITH, J.C. and MELTON, J.: Manipulation of Autopsy Diagnoses by Computer Techniques.
J. Amer. Med. Ass. 188 (1964) 958-962.

15) SMITH, J.C. and MELTON, J.: Automated Retrieval of Autopsy Diagnoses by Computer Technique.
Meth. Inf. Med. 2,3 p 85-9o (1963).

16) WOLFF-TERROINE? M., RIMBERT, D. and ROUALT, B.: Improved Statistical Methods for Automatic Construction of a Medical Thesaurus.
Meth. Inform. Med. 11 (1972) 1o4-112.

17) WOLFF-TERROINE, M., SIMON, N. and RIMBERT, D.: Use of a Computer for Compiling and Holding a Medical Thesaurus.
Meth. Inform. Med. 8 (1969) 34-4o.

18) WONG, R.L., GAYNON, P.: An automated parsing routine for diagnostic statements of surgical pathology reports.
Meth. Inform. Med. 1o: (1971) 168-175.

Konzeption und Organisation des AGK-Thesaurus

P. Röttger
F. Wingert
W. Feigl
P. Graepel
P. Ries
D. Schalk
W.M. Gross
F. Matakas

Universität Frankfurt a.Main
Senckenbergsches Zentrum der Pathologie
Direktor: Prof.Dr. W. Rotter

Medizinische Hochschule Hannover
Abteilung für medizinische Informatik
Direktor: Prof.Dr. P.L. Reichertz
Institut für Pathologie
Direktor: Prof.Dr. A. Georgii

Universität Wien
Pathologisches Institut
Vorstand: Prof.Dr. H.J. Holzner

Universität Bern
Pathologisches Institut
Direktor: Prof.Dr. H. Cottier

Rechenzentrum der Deutschen Klinik
für Diagnostik, Wiesbaden
Leiter: Dr.med. W. Giere

Klinikum Steglitz der Freien Universität Berlin
Pathologisches Institut
Direktor: Prof.Dr. W. Masshoff
Institut für Neuropathologie
Direktor: Prof.Dr. J. Cervós-Navarro

In Fortentwicklung des Experimentier-Thesaurus für Klartextanalyse, der vom Frankfurter Pathologischen Institut in Zusammenarbeit mit dem Deutschen Rechenzentrum in Darmstadt erarbeitet worden ist, wurde durch Mitarbeiter der Pathologischen Institute der Universität Berlin, Hannover, Frankfurt und Wien sowie des Institutes für Neurologische Pathologie in Berlin und der Deutschen Klinik für Diagnostik in Wiesbaden ein Thesaurus erstellt, dessen Organisation durch die Abteilung für Klinische Informatik der Medizinischen Hochschule Hannover durchgeführt wurde. Beratend wirkten bei diesem Konzept das Institut für Rechtsmedizin in Lübeck (PRIBILLA) sowie das Institut für Medizinische Dokumentation und Statistik in München (THURMAYR) mit.

In meinem folgenden Referat möchte ich vor allem die theoretische Konzeption des Systems darstellen und die Erörterung der Organisation auf einige Grundprinzipien beschränken; letztere soll an anderer Stelle unter Federführung von WINGERT in ausführlicher Form erfolgen.

1. Die allgemeine Konzeption des AGK-Thesaurus*)

Das System geht davon aus, daß medizinische Sachverhalte in der Form, in der sie in der zwischenärztlichen Routine-Kommunikation abgebildet werden, einer automatischen Auswertung zugänglich gemacht werden sollen. Die spezielle Form der Sachverhaltsdarstellung ist der jeweils zwischen Punkt und Punkt abgegrenzte Einzeldiagnosesatz (s. RÖTTGER et al., 1969). Die Struktur dieses Diagnosesatzes bestimmt das Bearbeitungs- und Auswertungs-System. Theoretisch sind zwei Satztypen denkbar: der einfachste Sachverhalt in Gestalt des Basis-Diagnosesatzes und der komplexe Sachverhalt in Gestalt des durch Lokalisation- und Modifikationsangaben erweiterten Diagnosesatzes.

1.1 Der Basis-Diagnosesatz

Um einen Sachverhalt in Worte abbilden zu können, stellt der Untersucher sich drei Grundfragen: Die Fragen "Was", "Wie" und "Wo". In Beantwortung dieser Grundfragen werden drei funktionell verschiedene Typen von Grundbegriffen verwendet. Die Antwort auf die Grundfrage "Was" erfolgt durch den Befundbegriff (das "finding"). Für diesen Begriff verwenden wir in der folgenden Darstellung das Symbol "F". Er stellt die Informationsbasis der Sachverhaltsmitteilung dar. Ein nur aus diesem Begriffstyp bestehender Elementardiagnosesatz würde beispielsweise nur aus dem Wort "Carcinom" bestehen. Formal liesse er sich dann durch das folgende Symbol darstellen:

. F . (1)

Ein derartig einfaches finding (wie der Befund "Carcinom") ist bereits als zwischenärztliche Mitteilung denkbar. Der Informationsgehalt dieser Nachricht ist sinnvoll, jedoch unzureichend.

---

*) Arbeitsgemeinschaft Klartextanalyse

Die findings werden in zwei Richtungen erweitert:

1) Durch die Beantwortung der Frage "Wie".

Für die modifizierten Angaben (modifiers) verwenden wir das Symbol "M". Unser Beispiel eines findings "Carcinom" würde durch den modifier "kleinzellig" in den erweiterten Diagnosesatz "kleinzelliges Carcinom" umgewandelt. Der Befund "F" kann in verschiedenen Modifikationen "M" vorliegen, somit läßt sich diese Erweiterung folgendermaßen darstellen:

. M (F) . (2)

2) Durch die Antwort auf die Frage "Wo".

Für die Lokalisationsbegriffe verwenden wir das Symbol "L". Der erweiterte Befund "M(F)" ist in verschiedenen Lokalisationen möglich, daher stellt er sich in allgemeiner Form folgendermaßen dar:

. L (M(F)) . (3)

Ein derartiger Basis-Diagnosesatz kommt in der zwischenärztlichen Kommunikation z.B. als scg. "Hauptdiagnose" zur Anwendung. Seinem Inhalt nach stellt er eine Grobklassifikation eines Sachverhaltes dar, die aus einer ganzen Reihe komplexer Befunde bzw. Sachverhalte herausgearbeitet sein kann.

1.2 Die Begriffsklassifikation

Die Einzeldiagnosesätze in Befundberichten bestehen aus Substantiven, Attributen und Partizipien. Diese Worte sind die Informationsträger, die in das Standardregister des Thesaurus aufgenommen werden. In das Begriffsklassifikations-System des AGK-Thesaurus wurde die Facetten-Konzeption des Frankfurt-Darmstädter Experimentier-Thesaurus mit ihrer auf zwei Stufen begrenzten Hierarchie für Lokalisations- und Befundbegriffe übernommen. Die F- und die L-Begriffe sind also jeweils in Ober- und Unterbegriffe unterteilt, wobei wir für die Unterbegriffe der Lokalisation das Symbol "l" und für Unterbegriffe der findings das Symbol "f" verwenden. Die modifiers werden nach wie vor nicht unterteilt.

Als neues Element hat der AGK-Thesaurus den "korrelierten Begriff" (correlative term) im Sinne der Facetten-Thesaurus-Konzeption von VICKERY aufgenommen.

Die Funktion eines korrelierten Begriffs läßt sich am zweckmäßigsten durch die Darstellung der verschiedenen Begriffstypen erläutern: Dabei gehen wir der Übersicht halber zunächst einmal davon aus, daß wir es bei den Begriffstypen von der Funktion her mit "reinen Worteinheiten" zu tun haben, daß

also nur Begriffe verwendet werden mit ausschließlicher F-, ausschließlicher M- und ausschließlicher L-Funktion.

1.2.1 Die Lokalisationsbegriffe

Der Lokalisationsbegriff "Koronararterie" hat die Zuordnung zu dem Oberbegriff "Arterie", als korrelierte Zuordnung den Oberbegriff "Herz". Das hat zur Folge, daß Befunde, die in der Lokalisation "Koronararterie" anfallen, sowohl im Begriffsfeld "Arterie" als auch im Begriffsfeld "Herz" registriert werden.

Ein anderes Beispiel für korrelierte übergeordnete Lokalisationen sind die Oberbegriffe "Gehirn" und "Rückenmark". Beide werden im AGK-Thesaurus nochmal einem weiteren Oberbegriff "Zentralnervensystem" zugeordnet. Die korrelierte Zuordnung "Gehirn" bezeichnen wir mit dem Symbol "L_c". Die übergeordnete" Lokalisation "ZNS" wird gekennzeichnet mit dem Symbol "L". Die Lokalisationsbenennung "Gehirn" läßt sich also für diese beiden Oberbegriffe mit dem Symbol

$$Lc(L) \qquad (4)$$

darstellen.

Diese Darstellung bedeutet, daß es sich bei der eingeengten Oberbegriffs-Lokalisation "Gehirn" um einen Bestandteil der Oberbegriffs-Lokalisation "ZNS" handelt.

Auf der Ebene der Unterbegriffe ergeben sich weitaus mehr Korrelationsmöglichkeiten. So können z.B. dem Unterbegriff "Bronchus" dessen weitere Einengungen "Oberlappenbronchus", "Mittellappenbronchus", "Unterlappenbronchus" usw. zugeordnet bzw. korreliert werden. Der korrelierte Unterbegriff modifiziert den übergeordneten Unterbegriff. In unserem Beispiel stellt "Oberlappenbronchus" eine Modifikation von "Bronchus" dar, was sich durch das Symbol

$$lc(l) \qquad (5)$$

darstellen läßt.

Selbstverständlich modifiziert ein Lokalisations-Unterbegriff grundsätzlich einen Lokalisations-Oberbegriff, so daß auch die Beziehung

$$l(L) \qquad (6)$$

gilt. Wenn wir für dieses System von korrelierten Ober- und Unterbegriffen uns eine bestimmte, in mehreren Dimensionen eingeengte Benennung vorstellen, so ergibt sich für sie - die komplexe Lokalisationsbenennung - das Symbol

$$lc(l(Lc(L))) \qquad (7)$$

1.2.2 Die Befundbenennung

Die gleiche Systematik ergibt sich bei der Abbildung=Benennung eines komplexen Befundes. Ein Unterbegriff wird einem Oberbegriff zugeordnet - der "Abszess" beispielsweise den lokalen Entzündungen -, was mit

. f(F) . (8)

ausgedrückt werden kann.

Bei Einführung korrelierter Befundbenennungen auf der Unterbegriffsebene ergibt sich die Darstellung

. fc(f) . (9)

und auf der Oberbegriffsebene

. Fc(F) . (10)

sowie in der gesamten Befundebene für den komplexen Befundbegriff

. fc(f(Fc(F)) . (11)

Die Unterbegriffsebene kann mit dem Begriff "Querfraktur" veranschaulicht werden, der dem gleichrangigen elementaren Befundbegriff "Fraktur" korreliert wird.

Die Oberbegriffsebene wird durch die Zuordnungsmöglichkeit des Befundes "Lungensilikose" veranschaulicht, die den Pneumokoniosen zugeordnet ist, wobei diese wiederum den Berufskrankheiten korreliert sind (denn alle Staublungenkrankheiten sind Berufskrankheiten, aber nicht alle Berufskrankheiten sind Staublungenkrankheiten).

1.2.3 Die modifier-Begriffe

Geht man davon aus, daß die echten modifier-Begriffe keine L- oder F-Funktion haben, so kann nach der Konzeption der erweiterten Basis-Information "M(F)" der modifier-Begriffstyp auch zur Modifikation der komplexen Befundbenennung verwendet werden. Damit ergibt sich für die modifizierte komplexe Befundbenennung folgende Darstellung:

. M(fc(f(Fc(F))) . (12)

1.2.4 Varia-Begriffe und insignifikante Begriffe

Analog den Verfahren im Experimentier-Thesaurus bestehen noch zwei weitere Klassifikationsmöglichkeiten. Einmal handelt es sich dabei um den Begriffstyp des insignifikanten Wortes. Es ist für die Erkennung eines encodierten Sachverhaltes ohne

Bedeutung und kann deswegen - nach seiner Erkennung durch das Thesaurus-System - aus dem Befundtext eliminiert werden. Zum anderen handelt es sich hier um den Varia-Begriffstyp. Diese Worte sind - gesehen von einer möglichen Textauswertung - von potentieller Bedeutung. Es kann nämlich von ihnen nicht ausgeschlossen werden, daß sie in irgendeiner logischen Beziehung die Basisbegriffe der Benennungen von Befunden, Lokalisationen und Modifikationen - kurz die F-, M- und L-Begriffe - modifizieren bzw. sogar in Frage stellen können. Vor allem betrifft dies Begriffe mit Negationinhalt. Aber auch Präpositionen, wie z.B. das Wort "mit" können bei einer Textaufschließung von Bedeutung werden (vgl. SCHALCK). Beispielsweise können durch diese Präpositionen zwei abgrenzbare Sachverhalte verknüpft sein, wobei sich als Lösung anbietet, den Diagnosesatz nicht als ungeeignet zu verwerfen, sondern die Präposition als Punkt zu lesen. Insgesamt ergeben sich auch durch die Varia-Begriffe lediglich formale Varianten von Diagnosesätzen, so daß wir sie in unserer theoretischen Darstellung eines komplexen Sachverhaltabbildes vernachlässigen können.

1.3 Der komplexe Diagnosesatz

Aus dem Basis-Diagnosesatz (3) und der komplexen Lokalisationsbenennung (7) sowie der modifizierten komplexen Befundbenennung (12) folgt für die Konfiguration des komplexen Diagnosesatzes die Darstellung:

. lc(l(Lc(L(M(fc(f(Fc(F))))))))) . (13)

Diese Formel besagt, daß mit den Darstellungsmöglichkeiten des AGK-Thesaurus die Projektion eines Sachverhaltes in 9 Variationsebenen bzw. Begriffsfeldern Verknüpfungen möglich sind. Den Kern der Information bildet der finding-Oberbegriff "F".

2. Die spezielle Konzeption des AGK-Thesaurus

Die implizierten Begriffsinhalte werden in der normalen zwischenärztlichen Kommunikation nicht mitgeteilt, ihre Kenntnis wird bei den Kommunikanten vorausgesetzt. Deswegen muß die automatische Standardisierung nicht nur die Worteinheiten selbst auf den gleichen formalen "Standard" bringen, sondern muß darüber hinaus auch die implizierten Begriffsinhalte in die Texte einbringen. In dieser Intention entspricht die Konzeption des AGK-Thesaurus der des Frankfurt-Darmstädter Experimentier-Thesaurus. Die Dimensionen, in denen Implikationen eingebracht werden, haben sich jedoch - wie das oben dargestellte Schema der möglichen Begriffsfeldebenen veranschaulicht - vor allem quantitativ erheblich erweitert.

Die Standardisierung der Primärtexte erfolgt über zwei nacheinander ablaufenden Verfahren, die wir mit Analyse und Synthese bezeichnen.

2.1 Die Analyse

Hierunter wird die einfache Implikationserschließung verstanden. Je nach Typ des zu analysierenden Wortes läuft sie unterschiedlich ab. Als wichtigste Unterteilung unterscheiden wir in den Befundtexten zwischen "Ein-Wort-Begriffen" und zwischen "Mehr-Wort-Begriffen".

2.1.1 Die Standardisierung der Ein-Wort-Begriffe

Die deutsche Sprache kennt eine Reihe von zusammengesetzten Worten, die sich in eine ganze Kette abgegrenzter Inhalte einschließen. An diesen Ein-Wort-Begriffen kann die Standardisierung komplett in einem Arbeitsgang vorgenommen werden, d.h. alle möglichen Zusatznotationen können in einer Arbeitsphase eingebracht werden.

2.1.2 Die Standardisierung von Mehr-Wort-Begriffen

Die Mehr-Wort-Begriffe bestehen aus Einzelworten, die erst zusammen - also im Kontext - einen eindeutigen Sinn ergeben und Oberbegriffen zugeordnet werden können. An den Bestandteilen dieser Mehr-Wort-Begriffe kann in der ersten Phase die Facetten-Zuordnung nur teilweise erfolgen. Zum Teil sind jedoch die Bestandteile eines Mehr-Wort-Begriffes identisch mit den Bestandteilen der Facetten-Notationskette eines Ein-Wort-Begriffes. Als einfachstes Beispiel hat der Ein-Wort-Begriff "Magencarcinom" die Notationskette "Magen" und "Carcinom"; der entsprechende Mehr-Wort-Begriff "Carcinom des Magens" ist nach Standardisierung von dieser Notationskette nicht zu unterscheiden. Nicht immer gibt es jedoch für Mehr-Wort-Begriffe einen zusammenfassenden Ein-Wort-Begriff. Dieser Umstand muß bei der weiteren Bearbeitung berücksichtigt werden.

2.1.3 Der Phasenablauf der analytischen Standardisierung

Dieser erste Standardisierungsvorgang läuft in 5 Phasen ab:
In der 1. Phase werden Schreibvarianten automatisch ausgeglichen, ohne daß den Texterstellern, den Kommunikanten, irgendwelche Schreibauflagen gemacht werden müssen.

In der 2. Phase werden aus dem Primärtext die Worte in ihrer ursprünglichen Form, die sog. Eingangsworte, in Standardbegriffe überführt. Die prefer terms werden automatisch gebildet

In der 3. Phase werden nach der Klassifikation zunächst die insignifikanten Begriffe nach ihrer Auffindung im Thesaurusregister aus den Texten gelöscht.

Die 4. Phase umfaßt die formale Klassifikation. Die Standardbegriffe werden als Ober- und Unterbegriffe der Lokalisation

bzw. der findings formal, d.h. durch eine Ziffer gekennzeichnet. Die primär modifizierenden Begriffe (M) sowie die Varia werden gleichfalls - jedoch ohne Unterteilung markiert.

In der 5. Phase erfolgt die inhaltliche Klassifikation, d.h. entsprechend dem Thesaurusregisterinhalt die Einbringung der Facetten-Notation. Für jeden komplexen Ein-Wort-Begriff wird die entsprechende Wortkette gebildet, so daß sich für diesen Textanteil bereits nach dieser ersten Standardisierungsphase die kompletten Begriffsfelder gebildet haben.

2.2 Die synthetische Standardisierung

Diese zweite Funktion des Thesaurus-Systems beinhaltet die komplexe Implikationserschließung. Sie erfolgt über die Erkennung von Wortketten in den Diagnosesätzen, wobei es gleichgültig ist, ob die Wortketten bereits in der primären Formulierung oder erst durch das Einbringen der Facetten-Notation gebildet worden sind. Die Konsequenz aus der "Erkennung" der Wortketten ist das Einbringen weiterer Facetten-Notationen, d.h. das Einbringen von superordinate terms zu entsprechenden Mehr-Wort-Begriffen in den Befundtexten. Die nach dieser Bearbeitungsphase vorliegende Form der Texte würde auch durch Überprüfung auf Übereinstimmung mit den kompletten Wortketten die Einbringung von Schlüsselziffern in die Diagnosesätze erlauben.

3. Organisationsprinzipien des AGK-Thesaurus

Der AGK-Thesaurus ist nach einem Doppel-file-System organisiert. Unterschieden wird das Eingangswortregister (E-Register) und das Standardwortregister (S-Register).

3.1 Das E-Register

Das Eingangswortregister enthält derzeit 26 245 Einheiten. Es konnte erheblich reduziert werden durch den automatischen Schreibweisenausgleich, auf den an anderer Stelle hingewiesen wird. Im Prinzip besteht dieser Schreibweisenausgleich darin, daß "Bindestriche" gelöscht werden, daß anstelle der möglichen Schreibweisen "c" oder "z", "c" oder "k" ausschließlich "c" gelesen wird und daß der Buchstabe "ß" grundsätzlich als "ss" erkannt wird, sowie, daß die Umlaute, die sich ja ohnehin nur in der deutschen Sprache befinden, einheitlich als Doppelbuchstaben gelesen werden.

3.2 Das S-Register

Das Eingangswortregister ist einem Standardwortregister zugeordnet. Dieses besteht aus 8473 Einheiten. Beide Register sind laufend numeriert. Die Identität einer Worteinheit wird bei der

E-Registerprüfung über die Übereinstimmung in der Buchstabenfolge festgestellt, bei allen weiteren Bearbeitungen über die Prüfung auf Übereinstimmung der Registernummer.

Die Textbearbeitung und die Textstandardisierung ist damit wesentlich vereinfacht worden. Das Facetten-Notation-Register ist so aufgebaut, daß jede Facetten-Notation sogleich Bestandteil des Standardwortregisters ist. Die Facetten-Notation kann also als Registernummer eingebracht und aufgesucht werden.

Insgesamt sind dem Standardregister 20 589 Facetten-Notationen hinzugefügt worden, wobei es sich um insgesamt 6387 verschiedene Notationen handelt. Die Größe der durch übergeordnete Befundbegriffe festgelegten Begriffsfelder schwankt zwischen 300 ("Degeneration/lokal") und 0 Standardworteinheiten, wobei die Tatsache, daß ein übergeordneter Begriff noch nicht zu Zuordnungen genutzt werden konnte, darauf zurückgeführt werden muß, daß dieser Thesaurus bisher in seinem Standardregister noch ein Ein-Wort-Begriffsregister ist. Der weitere Ausbau erfolgt empirisch. Wie der Frankfurt-Darmstädter Experimentier-Thesaurus ist das System "offen" konzipiert. Die jetzt in Frankfurt, Wien, Hannover und Berlin anlaufende oder angelaufene Textstandardisierung wird sowohl bei den Ein-Wort-Begriffen als auch bei den Mehr-Wort-Begriffen einer Reihe von Ergänzungen erbringen. Bezüglich der Einbringung der reinen Mehr-Wort-Begriffe ist auch der Vergleich mit einem Schlüsselsystem (WINGERT, GRAEPEL) für die Absteckung des Ergänzungsbedarfes vorgesehen.