Aus dem Institut für
Medizinische Biometrie und Medizinische Informatik
der Albert-Ludwigs-Universität Freiburg i.Br.

Morphembasierte automatische Indexierung und Klassifikation von Diagnosenbezeichnungen

Methodik, Realisierung und Evaluation

INAUGURAL-DISSERTATION
zur
Erlangung des Medizinischen Doktorgrades
der Medizinischen Fakultät
der Albert-Ludwigs-Universität Freiburg i.Br.

Vorgelegt Februar 2003
von Pius Franz
geboren in Bad Rippoldsau

Dekan:	**Prof. Dr. Martin Schumacher**
1. Gutachter:	**Prof. Dr. Rüdiger Klar**
2. Gutachter:	**Prof. Dr. Udo Hahn**
Jahr der Promotion:	**2003**

How to code an uncodified world?

No one has the answer.

But Medicine has been obliged to attempt it.

[Sager 95]

Vorwort

Eine der aktuellsten Herausforderungen der Medizin ist es, effiziente Mechanismen zum Erschließen einer immer größer werdenden Menge an medizinischer Information zu entwickeln. Der rapide Fortschritt der Computertechnologie einerseits, der eine ständig wachsende Fülle an Möglichkeiten der Informationsverarbeitung eröffnet, und die stetig zunehmenden Anforderungen an die Dokumentation und das Auffinden medizinischer Information andererseits stellen die Aufgabe, immer effektivere und den neuen Ansprüchen gerecht werdende Ansätze hervorzubringen.

Große Bereiche berühren sich dabei mit der Medizin. Konzepte der Dokumentations- und Ordnungslehre bilden die Grundlage für das Ordnen und Wiederfinden von Information. Die Informatik liefert Algorithmen und die Möglichkeiten zur Umsetzung von Such- und Codiersystemen. Voraussetzung für die Analyse von medizinischen Texten ist das weite Fachgebiet der Linguistik.

Diese Arbeit, die einen Beitrag zur Forschung im Bereich der Klassifikation von Diagnosen leisten will, stellt nach einer Einführung (Kapitel 1) und einem Abschnitt über wesentliche Grundlagen (Kapitel 2) in zwei methodischen Kapiteln (Kapitel 3 und 4) ein erweitertes klassisches Indexierverfahren und einen neuartigen Ansatz zur Klassifikation von Diagnosenbezeichnern dar, bevor nach einer Erläuterung der Realisation technischer Details (Kapitel 5) zwei Testverfahren und deren Ergebnisse vorgestellt werden (Kapitel 6), abgerundet durch deren Auswertung in einer abschließenden Diskussion (Kapitel 7). Ihr eigentliche Anliegen aber, wie aller Bemühungen in diesem Bereich, ist nicht ein wirkungsvoller Indexier- und Klassifikationsalgorithmus an sich, sondern durch ihn das verbesserte Verwalten und Erschließen medizinischer Information, die dem Arzt und letztendlich dem Patienten zu Gute kommen soll.

Inhalt

Publikationen

Teile dieser Arbeit wurden bereits publiziert oder sind in folgende Publikationen mit eingeflossen:

Franz P, Zaiß A, Schulz S, Hahn U, Klar R (2000). *Automated Coding of Diagnoses – Three Methods Compared.* Proceedings of the 2000 AMIA Annual Fall Symposium, 250-254

Schulz S, Romacker M, Franz P, Zaiß A, Klar R, Hahn U (1999). *Towards a multilingual morpheme thesaurus for medical free-text retrieval.* Medical Informatics Europe 1999, 891-894

Schulz S, Hahn U (2000). *Morpheme-Based, Cross-Lingual Indexing for medical Document Retrieval.* International Journal of Medical Informatics 2000; 58-59: 87-99

1. Einführung

Klassifikationen von Krankheiten und ihren Folgen, von Medikamenten, Operationen und anderen therapeutischen oder diagnostischen Verfahren unterstützen die begriffliche Ordnung des medizinischen Wissens. Damit dienen sie unter anderem behandelnden und forschenden Ärzten, den medizinischen Dienstleistern und Kostenträgern sowie öffentlichen Stellen bei der standardisierten medizinischen Dokumentation und deren Auswertung zu Zwecken der Krankenhausstatistik, Epidemiologie, Kostenplanung und mehr. Eine korrekte und effiziente medizinische Verschlüsselung ist heute für jedes Krankenhaus fundamental, da aus der medizinischen Basisdokumentation wesentliche Daten für die Abrechnung, die Pflegesatzverhandlungen und für das interne Controlling und Qualitätsmanagement abgeleitet werden. Insbesondere die Nutzung als Leistungsdokumentation und ihre Verknüpfung mit der Finanzierung des Gesundheitswesens im Rahmen von *Diagnosis Related Groups* (DRGs) gewinnt aktuell immer mehr an Bedeutung und setzt das Problem der korrekten Codierung an die Spitze der derzeitigen Forschungsbemühungen im Bereich der Krankenhausadministration [Lauterbach 00, GMDS 97, Graubner 95].

1.1 Gesetzliche Regelungen über die medizinische Dokumentation

Die Anwendung von Standards bei der medizinischen Dokumentation ist Voraussetzung für eine hohe Qualität der Datensammlung, während die Nutzung medizinischer Klassifikationen die Vergleichbarkeit dieser Daten und damit ihre Interpretationsmöglichkeit unterstützt [Diekmann 92]. Beispiele für solche Standards im Krankenhausbereich, die auf gesetzlichen Grundlagen beruhen, waren in der DDR die seit 1968 bzw. 1979 geltenden Verordnungen über den Krankenblattsignierstreifen, mit dem Diagnosen nach der Internationalen Klassifikation der Krankheiten, Verletzungen und Todesursachen, 8. bzw. später 9. Revision (ICD-8 bzw. ICD-9) vierstellig verschlüsselt wurden. Die Krankenhäuser der alten Bundesrepublik, die dem Krankenhausfinanzierungsgesetz unterliegen, wurden 1986 im Rahmen der Bundespflegesatzverordnung (BPflV) von 1985 erstmals verpflichtet, für die Krankenkassen eine abteilungsbezogene Diagnosenstatistik zu erstellen und dafür in einer kurzen Basisdokumentation die Entlassungsdiagnosen aller stationären Behandlungsfälle nach der ICD-9 dreistellig zu codieren. – Mit der Krankenhausstatistik-Verordnung von 1990 wurde ab 1993 die landesweite anonymisierte Zusammenfassung dieser Daten festgelegt, die es früher schon, und zwar personenbezogen, in der DDR gegeben hatte.

Das Gesundheitsstrukturgesetz (GSG) von 1992 hat mit der Änderung des Fünften Buches des Sozialgesetzbuchs (SGB V) den Datenumfang wesentlich ausgeweitet, um vor allem den Kostenträgern eine genauere Überprüfung der Krankenhausbehandlungen zu ermöglichen. Durch § 295 SGB V werden die Mediziner der kassenärztlichen Vereinigung verpflichtet, Diagnosen nach ICD zu codieren. Entsprechend den §§ 301 und 303 SGB V fordert der Gesetzgeber seit 1995 die verschlüsselte und maschinenlesbare Übermittlung von Einweisungs-, Aufnahme-, Verlaufs-, Entlassungs- und Verlegungsdiagnosen bei externer Verlegung (abgebildet mit der ICD) sowie von Operationen (definiert mit dem OP-Schlüssel gemäß § 301 SGB V) für die Abrechnung mit den Krankenkassen [Sozialgesetzbuch 96]. Gleichzeitig wurde aus Transparenzgründen auch für den vertragsärztlichen Bereich (ambulantes Gesundheitswesen) die Diagnosendokumentation nach den ICD-Subkategorien auf den Abrechnungsunterlagen und Arbeitsunfähigkeitsbescheinigungen festgelegt. Diese seit 1996 geltenden gesetzlichen Regelungen verwirklichen in Deutschland weitgehend den Vorschlag für ein *Minimum Basic Data Set* (MBDS) der damaligen Europäischen Gemeinschaft (EG) von 1982, der als medizinische Merkmale für die medizinische Basisdokumentation stationärer Behandlungsfälle Diagnosen sowie diagnostische und therapeutische Prozeduren vorsieht.

Die BPflV von 1995 schreibt für die Krankenhäuser eine ICD-Diagnosen- und eine OPS-301-Operationsstatistik vor. Weiter sind die Fallpauschalen und Sonderentgelte mit den in der BPflV festgelegten ICD- und OPS-301-Codes zu übermitteln. – Durch das GSG von 1992 sowie die BPflV von 1995 werden Diagnosen und Operationen, definiert mit der ICD und der deutschen Fassung der Internationalen Klassifikation der Prozeduren in der Medizin (ICPM-GE) bzw. dem damit kompatiblen Operationsschlüssel OPS-301, zu zentralen Leitindikatoren für die Leistungsbewertung und Finanzierung des Gesundheitswesens und damit unter anderem zur Grundlage für künftige Pflegesatzverhandlungen. Die Diagnosenverschlüsselung erfolgt darüber hinaus auch für Zwecke der wissenschaftlichen Auswertbarkeit einschließlich epidemiologischer Fragestellungen und des Qualitätsmanagements [GMDS 96]. Über einen diagnoseorientierten Leistungs- und Aufwandsvergleich zwischen den Krankenhäusern der Bundesrepublik Deutschland sollen Möglichkeiten zur Kostendämpfung im Gesundheitswesen aufgezeigt werden.

Hieraus folgt die Notwendigkeit einer standardisierten Definition, Registrierung und Codierung der Diagnosen, die eine valide und reproduzierbare Umsetzung medizinischer Sachverhalte ermöglichen [GMDS 97, Diekmann 92]. Mit der 10. Revision der ICD (ICD-10) wurde

1998 im Krankenhausbereich nun eine moderne, völlig neue Systematik für die Diagnosencodierung eingeführt, in der sich der Fortschritt der Medizin widerspiegelt, die aber wie jede Weiterentwicklung der gesetzlichen Richtlinien bisher zu einer zusätzlichen Mitarbeiterbelastung führte.

Die Anforderungen an die Dokumentation des medizinischen Leistungsgeschehens werden sich durch die ab 2003/2004 geplante Umstellung der Vergütungsregelung für Krankenhäuser auf das System der *Diagnosis Related Groups* (DRGs) auch in den nächsten Jahren erheblich erweitern, da dann nahezu die gesamte Finanzierung des Krankenhauses auf der Grundlage der Daten der Basisdokumentation, insbesondere der codierten Diagnosen und Prozeduren, erfolgt [Lauterbach 00]. Ähnlich wie bisher bereits bei Fallpauschalen und Sonderentgelten orientiert sich dabei die Finanzierung im wesentlichen nicht an der Verweildauer, sondern innerhalb gewisser Grenzzeiträume ausschließlich an der Diagnose und den zugehörigen Kostengewichten. Die eigentliche Abrechnung der erbrachten Leistung schreibt nun der Arzt mit der Codierung. – Dieses System birgt nach den Erfahrungen, die seit seiner Einführung in Frankreich und Australien gemacht wurden, ein deutliches Potenzial an Budgeteinsparungen in sich. Um eine adäquate Abbildung des Leistungssgeschehens mittels DRGs zu erreichen, die nun die Basis für die Vergütung der durchgeführten Behandlung eines Patienten darstellen und die Rechnungslegung des Krankenhauses bestimmen, ist allerdings eine lückenlose und qualitativ hochwertige Codierung von Diagnosen und Prozeduren notwendig [Roeder 02]. Dabei gewinnt auch die Codierung von Prozeduren und Nebendiagnosen, von denen bis zu 5 in die fallgruppenbezogene Abrechnung mit eingehen, erheblich an Bedeutung [Metzger 02]. Nichtärztliche Berufsgruppen, zum Beispiel die Pflege, tragen durch zusätzliche Informationen wesentlich zur genauen Darstellung eines Behandlungsfalles bei; eine vorzügliche Codierung sollte daher auch von wenig mit Kodierrichtlinien vertrautem Personal zumindest in medizinischen Teilbereichen möglich sein [Stiller 02]. Wesentlicher Nachteil der künftigen Nutzung der Diagnosencodierung zu Abrechnungszwecken ist, dass die Dokumentation immer mehr nach administrativen Gesichtspunkten ausgerichtet wird und ursprünglich auf den Patienten zentrierte Zielsetzungen, wie das Erlangen klinisch und epidemiologisch relevanter Informationen durch Mittel der Statistik, nachrangig werden.

Eine weitere gesetzliche Regelung, welche die ärztliche Diagnose direkt mit finanziellen Leistungen in Verbindung bringt, ist die Einführung eines morbiditätsbezogenen Risikostrukturausgleichs ab spätestens 2007. – Da seit 1996 die Bürger ihre Krankenkasse selbst wählen können, sind die seit 1994 vorgeschriebenen Ausgleichszahlungen zwischen den einzelnen

Kassen erforderlich, um unterschiedliche Risiken bei der historisch gewachsenen Mitgliederverteilung finanziell anzupassen. Dieser sogenannte Risikostrukturausgleich berücksichtigte jedoch die tatsächliche Morbiditätsverteilung nicht, so dass sich die Kassen in der Folge hauptsächlich um junge und gesunde Kunden bemühten. Um diesem Dilemma abzuhelfen, das momentan durch Übergangsregelungen überbrückt wird, ist ein morbiditätsorientierter Risikostrukturausgleich durch eine direkte Erfassung des Gesundheitszustandes der Versicherten notwendig; die Ausgleichszahlungen erfolgen dann entsprechend den nach der ICD verschlüsselten Diagnosen und den tatsächlich anfallenden Behandlungskosten für Patienten.

Eine ausreichende Güte der Codierung von Diagnosen ist aus all diesen Gründen heraus dringend notwendig. Sie kann ohne weitere Belastung des medizinischen Personals durch administrative Aufgaben aber nur gewährleistet werden, wenn in zunehmendem Maße DV-gestützte Systeme zur Indexierung und Klassifikation zur Verfügung stehen, wie u.a. von Fachgremien wie der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS) oder der europäische Standardisierungsinstitution (CEN) und den Krankenkassen gefordert wird.

1.2 Güte aktueller klinischer Methoden der Diagnosencodierung

Gib jeder Sache ihren ganz bestimmten Platz und stelle sie dann woanders hin.
Dies ist zwar kein weiser Rat, dafür aber eine verbreitete Sitte. (Mark Twain)

Diagnosen und Operationen haben in der Medizin als Texte definiert die höchste Aussagekraft. Die Projektion von Klartexten auf ein ein- oder mehrdimensionales Verschlüsselungssystem ist in aller Regel mit einem Informationsverlust verbunden [GMDS 96], idealerweise (in der Praxis leider oft nicht) ein Verlust von Fakten, die für die Zielsetzung der Klassifikation irrelevant sind. Ebenso kommt es zu Fehlzuordnungen in Fällen, in denen grundsätzliche Probleme der ICD bzw. des OPS-301 es erschweren, eindeutige Codes für Diagnosentexte oder Prozeduren zu finden.

Zurzeit werden drei Grundtypen von Verschlüsselungsverfahren eingesetzt, nämlich

1. die manuelle Codierung anhand des ICD-10- bzw. OPS-301-Katalogs durch den Arzt auf Station,

2. die Erfassung über vorcodierte Erhebungsbögen und

3. die Zuordnung über DV-gestützte Klassifikationssysteme.

1. - In einer Studie zur Qualität verschiedener Verfahren der Diagnoseverschlüsselung am Universitätsklinikum Tübingen [Bosing 96] wurde das bei weitem schlechteste Ergebnis bei der manuellen Codierung durch den Arzt auf Station erzielt mit einer Fehlerhäufigkeit von bis zu 36,8 %. Ähnliche Fehlerquoten sind auch aus anderen Kliniken bekannt; so ergab zum Beispiel die Nachverschlüsselung von 1221 Diagnosen in Bochum eine Fehlerrate von etwa 38 % bei der Codierung von Hauptdiagnosen [Nietzschke 92]. Dabei schneidet die manuelle Verschlüsselung mit Hilfe eines Browsers nicht besser ab [Nilsson 00]. Die Ursachen für einen derartig hohen Qualitätsverlust dürften u.a. in der steigenden Belastung der Mediziner durch administrative Pflichten liegen, deren Nutzen für viele nicht einsehbar ist. – Auch wo die Codierung durch dafür speziell ausgebildete medizinische Dokumentare vorgenommen wird, ist die Häufigkeit von inkorrekten Verschlüsselungen noch erstaunlich hoch. So beträgt zum Beispiel in Australien, wo die Abrechnung über ein DRG-System, an dem sich das deutsche orientiert [Schlottmann 02], direkt an die Verschlüsselung gekoppelt ist, die Fehlerquote noch immer 13 %, bei einem finanziellen Verlust von allerdings nur ca. 1 %, was für Fehlklassifikationen aus Unkenntnis und nicht aus Vorsatz spricht.

2. - Bei der zweiten Verschlüsselungsmethode werden durch Textstandardisierungen (zum Beispiel vorcodierte Erhebungsbögen) bereits Verbesserungen gegenüber der manuellen Diagnosecodierung durch den Arzt auf Station erreicht. So kann etwa durch Top-Listen die Verschlüsselung der häufigsten abteilungsbezogenen Haupt- und Nebendiagnosen erleichtert werden [Roeder 02]. Eine exakte Untersuchung, inwieweit die auf den Menübögen markierten Diagnosen mit den tatsächlichen Angaben in der Krankenakte übereinstimmen, steht allerdings noch aus.

3. - Im Vergleich dazu führt die automatische und semiautomatische (Auswahl aus einer Liste vom System vorgeschlagener Klassen) Verschlüsselung zu signifikant besseren Resultaten [Bosing 96]. Die Anzahl an inkompletten Codierungen sinkt, während sich die Anzahl korrekter Klassifikationen erheblich steigert; darüber hinaus reduziert sich der Zeitaufwand auf rund

die Hälfte [Hohnloser 96]. Dennoch bietet der Einsatz eines automatischen Codierverfahrens allein noch keine Gewähr für eine optimale Verschlüsselung. Unter anderem können Fehlcodierungen durch systematische Auswirkung eine erhebliche Verzerrung des wahren Krankheitsspektrums nach sich ziehen. – DV-unterstützte Codiersysteme wie Diacos 6.0, Kodip oder ICD-ICPM professional sind heute in allen größeren Krankenhäusern der Bundesrepublik gängig.

Der Einsatz von Computern erleichtert die Klassifikation von Diagnosen und Prozeduren und verbessert ihre Güte. Dabei stellen die Eigenschaften von Sprache im Allgemeinen (z.B. Ambiguitäten, Synonyme, Periphrasen) und von medizinischer Fachsprache im Besonderen (z.B. großes Vokabular, komprimierter Telegrammstil, Akronyme) Herausforderungen, die zu einer Reihe von Ansätzen für entsprechende Algorithmen geführt haben. Grundzüge solcher Verfahren sollen nun im Folgenden vorgestellt werden.

1.3 Grundlegende Verfahren zur DV-gestützten Klassifikation

Die automatische oder semiautomatische Klassifikation von Suchanfragen (*Queries*) erfordert in einem ersten Schritt das sogenannte *Indexieren*, bei dem jedem Dokument des Suchraums (Dokumentenkollektion, *Retrieval Space*) sowie in der Folge auch allen Suchanfragen ein eindeutiger Dokumentbezeichner (Deskriptor, *Identifier*) zugeordnet wird. In einem zweiten Schritt, dem Wiederfinden von Dokumenten (*Retrieval*), wird dem Bezeichner einer Suchanfrage ein oder eine Menge von Dokumenten des Suchraums (bzw. deren Deskriptoren) zugeordnet, gegebenenfalls mit Gewichtung. Bei der semiautomatischen Klassifikation wählt der Benutzer anschließend das gesuchte Dokument aus, das in der Regel in der sich ergebenden Teilmenge der Dokumentenkollektion liegen wird; bei der vollautomatischen Klassifikation trifft das System diese Entscheidung, in der Regel nach topologischen (d.h. den Abstand von Dokumenten betreffenden) oder probabilistischen Gesichtspunkten. – Für eine ausführlichere Erläuterung der hier eingeführten Begriffe des Information Retrieval in bezug auf die Medizin siehe [Klar 97].

Maschinelles Indexieren von Dokumenten

Da Dokumente (im weiteren Sinne sollen auch Suchanfragen als Dokumente angesehen werden, obwohl dieser Begriff primär Elemente der Dokumentenkollektion bezeichnet) im Fol-

genden ausschließlich aus Text bestehen sollen, kann sich ein Deskriptor im einfachsten Fall aus der Zeichenkette des Dokumentes zusammensetzen, oder etwa aus dem Vektor der im Dokument enthaltenen Wörter in der Reihenfolge ihres Auftretens. Dabei sind Wörter definiert als Zeichenketten, die durch bestimmte Sonderzeichen abgegrenzt sind. – Die Wahl des Bezeichners beeinflusst wesentlich, welche Elemente des Dokumentenraumes im späteren Retrieval einander ähnlich sind, d.h. sich im Sinne einer Metrik nahe stehen. Jedem Dokument des Suchraumes muss ein Bezeichner zugeordnet sein, der innerhalb des Suchraumes ausschließlich dieses Element repräsentiert, da ein Rückschluss vom Bezeichner auf das Dokument möglich sein muss (Injektivität). Prinzipiell können Dokumente außerhalb des Suchraums (also Suchanfragen) jedoch durchaus auf gleiche Deskriptoren projiziert werden. So würde zum Beispiel die Projektion eines Dokumenttextes auf den gleichen Text in Kleinbuchstaben viele Schreibvarianten vereinheitlichen, reduziert auf den Suchraum z.B. der ICD-Diagnosenbezeichner wäre eine Rückführung auf das Herkunftselement aber weiterhin möglich.

Eine manuelle Indexierung von Dokumenten erfolgt generell nur für ausgewählte Texte. Die ICD-Schlüssel sowohl des Suchraumes als auch des Testdatensatzes an klinischen Entlassungsdiagnosen, den wir verwenden, wurden manuell attribuiert. Da Ziel dieser Arbeit die maschinelle Klassifikation von Diagnosen ist, kommt ein manuelles Zuordnen von Bezeichnern zu Dokumenten innerhalb unseres Algorithmus per Definition nicht in Frage.

Die einfachsten automatischen Indexierverfahren orientieren sich an der **Zeichenkette** des zu indexierenden Dokumentes, wobei unter Zeichen Buchstaben, Ziffern und Sonderzeichen (z.B. das Leerzeichen) zusammengefasst sind. Bei dieser Art der Indexierung wird auf ein inhaltliches Erfassen des Textes vollständig verzichtet. Technisch einfach ist ebenfalls die Projektion eines Textwortes auf eine Vorzugsbenennung. So können auch auf der Ebene der Zeichenketten bereits Synonyme erschlossen oder Abkürzungen aufgelöst werden, wenn diese in vorab definierten Wörterbüchern enthalten sind. Auch die Schreibweise kann durch die Abbildung von Textwörtern auf normierte Zeichenketten (z.B. Projektion auf Kleinbuchstaben, Ersetzen von ß durch ss usw.) uniformisiert werden.

Ebenfalls einzig auf Grundlage der Zeichenketten basiert das klassische Verfahren der **N-gramm-Indexierung,** bei der Dokumente durch Vektoren repräsentiert werden, die aus allen möglichen Zeichenketten der Länge N bestehen, die im Text des Dokuments enthalten sind. So würde etwa die Zeichenkette $Z_1...Z_S$ durch den Deskriptor ($Z_1...Z_N$, $Z_2...Z_{N+1}$, ...,

$Z_{S-N+1}...Z_S$) indexiert. Der Sonderfall der Trigramm-Indexierung (N=3) wurde zum experimentellen Vergleich unseres Algorithmus mit einem herkömmlichen Verfahren hergezogen, er wird daher in Kapitel 6.2 näher erläutert.

Ein etwas anspruchsvollerer Ansatz liegt dem **linguistisch basierten Indexieren** zu Grunde. Dieser Ansatz beruht auf morphologischer Analyse (Morphologie = Wortaufbau), grammatischer bzw. syntaktischer Verarbeitung (Syntax = Satzbau) und semantischer Zuordnung (Semantik = Lehre von den Wortbedeutungen). Dabei können auch nur Teile dieser Methoden zum Einsatz kommen.

Bei der morphologischen Analyse eines Textworts wird versucht, dieses in Wortteile zu zerlegen. Atomare, d.h. nicht weiter zerlegbare Teile werden dabei als Morpheme bezeichnet. Zu diesen Morphemen zählen Wortstämme, aber auch z.B. Endungen, also Zeichenketten mit sehr unterschiedlichem Informationsgehalt. Bei der morphologischen Analyse kommen unterschiedlich komplexe Verfahren zum Einsatz. Einfachstes Beispiel wäre das sogenannte *Stemming*, bei dem durch Abspalten von Endungen Wortstämme identifiziert werden können.

Die anschließende grammatikalische Analyse versucht, einzelne Satzteile (z.B. Subjekt, Prädikat, Objekt...) zu erkennen. Obwohl der Aufbau von Freitextdiagnosen oft keine kompletten Sätze widerspiegelt, kann eine syntaktische Zerlegung von großem Nutzen sein, etwa bei der Abgrenzung von Haupt- und Nebendiagnosen. – Wenn die einzelnen Wortteile analysiert worden sind und ggf. das Dokument syntaktisch zerlegt wurde, ist es relativ einfach, aus den wichtigsten Morphemen Deskriptoren abzuleiten und so eine semantische Zuordnung festzulegen.

Für einen guten Überblick über die Medizinische Linguistik und ihre grundlegenden Begriffe sei verwiesen auf [Ingenerf 97].

Eine Erweiterungsmöglichkeit des linguistisch basierten Indexierens wäre das **wissensbasierte Indexieren.** Dabei wird durch Darstellung von Äquivalenzklassen, Gegenteil von-Beziehungen, hierarchischen Beziehungen usw. ein semantisches Netzwerk erstellt, mit dem Verbindungen zu Textteilen mit bereits erkennbarem Inhalt des Dokuments geknüpft werden können. So ist es möglich, durch Hintergrundwissen aus dem Sachgebiet des Dokuments, etwa aus sogenannten Expertensystemen, die semantische Zuordnung zu verbessern und Mehrdeutigkeiten aufzulösen, wie etwa bei Homonymen wie „Blase“ oder bei Abkürzungen

wie „syst.“ (systolisch? systemisch? systematisch?). Die meisten wissensbasierten Indexierverfahren sind allerdings noch im Stadium intensiver Forschung.

Eine ausführlichere Beschreibung aktueller Indexiermethoden bietet [Gaus 00].

Grundlegende Retrievalverfahren

Die eigentliche Suche eines (nächstgelegenen) Dokuments in einer Dokumentenkollektion, das Retrieval, erfolgt nun auf Basis der durch die Indexierung ermittelten Deskriptoren. Dabei ist die Retrievalqualität sowohl abhängig von der Qualität des Index als auch von der Qualität des eigentlichen Retrievalalgorithmus, sowie davon, wie Index und Algorithmus aufeinander abgestimmt sind. In der Regel gibt es daher zu jedem Indexierverfahren ein Retrieval, das sich dem Konzept des Index relativ natürlich anfügen lässt, weshalb Indexierung und Suche in vielen Publikationen in einem Schritt dargestellt werden. Es ist aber prinzipiell möglich, sowohl auf den gleichen Index verschiedene Retrievalmethoden anzuwenden, als auch ein spezielles Retrieval auf unterschiedliche Deskriptorensammlungen, eine Tatsache, die sich unter anderem zu Testzwecken als nützlich erweist.

Die einfachste Art des Retrievals, die sich sinnvoll an die zeichenkettenorientierte Indexierung anfügen lässt, ist die **Boolesche Suche.** Bei diesem Verfahren werden mit Hilfe von Booleschen Operatoren (logisches Und, Oder, Nicht) Deskriptoren der Anfrage auf Übereinstimmung mit Bezeichnern von Dokumenten des Suchraums überprüft. Zu diesem klassischen Verfahren gibt es eine Reihe von Erweiterungen, wie etwa den NEAR-Operator, bei dem Terme der Anfrage nicht nur gleichzeitig in Dokumenten enthalten sein müssen, sondern auch in räumlicher Nähe. Die Terme können dabei aus den im Text vorkommenden Wörtern bestehen, prinzipiell aber auch aus allen anderen Indizes wie etwa Trigrammen. Nachteil des Booleschen Retrievals ist vor allem die schlechte Retrievalqualität. Die Antwortmenge ist nicht nach mehr oder weniger relevanten Dokumenten unterteilt, und die Trennung zwischen gefundenen und nicht gefundenen Dokumenten sehr streng, da etwa bei der Anfrage $T_1 \wedge T_2 \wedge T_3$ alle Dokumente, die zwei der Terme enthalten, genauso verworfen werden wie solche, die mit der Menge $\{T_1, T_2, T_3\}$ völlig disjunkt sind. – Eine Erweiterung des Booleschen Retrievals, die dessen Nachteile allerdings nur teilweise überwindet, wäre das **Fuzzy Retrieval,** bei dem eine gewichtete Indexierung stattfindet und dadurch ein Ranking der Ergebnismenge möglich ist. Sinnvollen Einsatz findet die exakte Zeichenkettensuche beim Durchsuchen von

Dokumenten, die nicht vorindexiert sind, weil sie z.B. als reine Textdateien über Internetquellen wie MEDLINE bezogen wurden [Lovis 00].

Einen etwas komplexeren Ansatz als das Boolesche Retrieval verfolgt das **Vektorraummodell**, bei dem Dokumente und Anfragen als Punkte in dem orthonormalen Vektorraum aufgefasst werden, der durch die Deskriptoren der Elemente des Suchraums aufgespannt wird. Dadurch können Ähnlichkeitsmaße wie z.B. das Cosinusmaß, Gewichtungsverfahren und ein *Relevance Feedback* zur Verbesserung der Termgewichte bei künftigen Anfragen realisiert werden. Dieses klassische Modell, das auf Salton zurückgeht, spielt in der Dokumentensuche nach wie vor eine wichtige Rolle, da die Retrievalqualität im Vergleich zum Booleschen Ansatz deutlich höher ist. Das Vektorraum-Retrieval wurde wie die Trigrammindexierung zum Vergleich mit unserem Verfahren herangezogen, es wird daher in 6.2 noch genauer dargestellt.

Eine Erweiterung des Vektorraummodells kann durch das **Dokumenten-Clustering** geschehen, bei dem davon ausgegangen wird, dass die Ähnlichkeit relevanter bzw. irrelevanter Dokumente untereinander größer ist als zwischen zufälligen Teilmengen der Dokumentenkollektion. So könnten etwa alle ICD-Klassen, unter die Karzinome fallen, zu einem Cluster zusammengefasst werden. Die Suche nach einer Anfrage, die den Term „Karzinom“ enthält, kann nun innerhalb dieses Clusters schneller und präziser erfolgen. Durch die Unsicherheitsfaktoren, die die Zuordnung einer gegebenenfalls umfangreichen Anfrage zu solchen Clustern aufgrund von Ähnlichkeiten mit sich bringt, ist die Retrievalqualität beim Dokumenten-Clustering jedoch im Allgemeinen niedriger als die des reinen Vektorraum-Retrievals.

Weitere Möglichkeiten des Retrievals bieten **probabilistische Verfahren,** die aus Lerndatensätzen oder früheren Anfragen lernen, sowie **heuristische Methoden,** die auf Hintergrundwissen aus dem Sachgebiet des Dokuments oder speziellen Eigenschaften des zu Grunde liegenden Index beruhen. Ein rasches und einfaches probabilistisches Verfahren zum Dokumenten-Retrieval bietet etwa [Hersh 95]. Eine umfassendere Zusammenstellung derzeitiger Retrievalalgorithmen findet sich in [Rechenberg 02]; einen guten Überblick über die wichtigsten Verfahren bietet [Wiesman 97]. Für eine ausführliche Darstellung des umfangreichen Gebiets des Information-Retrieval in der Medizin sei auf [Hersh 96] verwiesen.

Für den Vergleich von Retrievalmethoden eignen sich im Allgemeinen die Qualitätsmaße ***Precision*** (Präzision) und ***Recall*** (Abdeckung). Dabei bezeichnet der *Recall* eines Verfahrens

den Anteil der relevanten Dokumente, die mit einer bestimmten Anfrage gefunden wurden, bezogen auf die (meist schwer zu ermittelnde) Gesamtzahl aller relevanten Dokumente im Suchraum. Die Präzision hingegen gibt den Anteil relevanter Dokumente an der – bekannten – Ergebnismenge der Anfrage an; sie ist abhängig von der Prävalenz, dem Anteil relevanter Dokumente im Suchraum als solchem. Da die Bestimmung von Recall und Präzision nur hilfreich sind, wenn die Ergebnismenge von Anfragen aus mehr als einem Element bestehen, konnten sie zur Beurteilung unseres Verfahrens allerdings nicht zu Rate gezogen werden.

Die Entwicklung neuer Retrievalalgorithmen bildet nicht nur bezogen auf die Medizin ein großes und spannendes Arbeitsfeld. Viele der in den Krankenhäusern aktuell eingesetzten Verfahren beschränken sich weitgehend darauf, Klassifikationen und große Lerndatensätze auf der Suche nach rein syntaktischer Ähnlichkeiten zu browsen, ohne auf morphologischen Aufbau oder gar semantische Bedeutung der Suchanfragen einzugehen. Um die Grenzen solcher Systeme zu überwinden, sind neue Ansätze notwendig, die verstärkt auf den Inhalt von Diagnosentexten eingehen und ihn linguistisch analysieren.

Im Folgenden werden die Weiterentwicklung eines morphembasierten automatischen Indexierverfahrens (Kapitel 3) und der Entwurf einer neuen Retrievalmethode (Kapitel 4) ausführlich beschrieben. Zuvor aber schließt sich ein Kapitel über die begrifflichen Ordnungssysteme in der Medizin an, um die es letztlich geht.

2 Grundlagen zu begrifflichen Ordnungssystemen in der Medizin

2.1 Klassifikationen und Nomenklaturen

Wie im vorigen Kapitel bereits angesprochen wurde, gibt es für die Verschlüsselung von Diagnosen verschiedene Konzepte. Als wichtigste Varianten haben sich dabei die der Klassifikation und die der Nomenklatur herausgestellt, die sich grundlegend unterscheiden. Da sich eine Codierung immer auf eine spätere Auswertung bezieht, leiten sich unterschiedliche Ansätze stets aus unterschiedlichen Fragestellungen ab, die an ein Ordnungssystem gerichtet werden. Daher haben Klassifikationen und Nomenklaturen zum Teil sehr verschiedene Einsatzgebiete.

Klassifikationen haben die Aufgabe, Dokumente anhand des bezogen auf die Fragestellung wesentlichen, klassenbildenden Merkmals zusammenzufassen. In unserem Falle etwa sollen Krankheitsbezeichnungen hinsichtlich der Frage nach der Hauptdiagnose einsortiert werden. Klassifikationen sind meist monoaxial und ordnen zu verschlüsselnde Dokumente je genau einer Klasse zu, weshalb sie sich in erster Linie für statistische Auswertungen eignen. Einer Unvollständigkeit des Ordnungssystems wird dadurch vorgebeugt, dass eine oder mehrere Klassen für anderweitig nicht klassifizierbare Elemente bereitgestellt werden. Da bei der Klassifikation eines Dokuments alle sekundären Merkmale und Informationen verworfen werden, ist eine Abbildung von der Klasse bzw. deren Schlüssel (Code) auf die ursprüngliche Krankheitsbezeichnung nicht mehr möglich. Dieser Informationsverlust ist insbesondere dann gravierend, wenn die ursprüngliche Fragestellung, anhand derer klassifiziert wurde, durch eine andere ersetzt werden soll. Für ein Retrieval, das auf unterschiedliche Anfragen flexibel reagieren können muss, sind Klassifikationen daher nur sehr begrenzt geeignet.

Nomenklaturen zeichnen sich dadurch aus, dass im Prinzip sämtliche bezogen auf eine Grundgesamtheit (bei uns die Medizin) relevanten Informationen codiert werden. Ein Dokument wird damit einem Vektor zugeordnet, der aus Deskriptoren für die verschiedenen Informationseinheiten des Dokuments besteht, die im vorgegebenen Ordnungssystem enthalten sind (Indexierung). Diese Art der Verschlüsselung trifft keine Aussage über die semantische Gewichtung der verschiedenen Informationseinheiten. Es ist daher zum Beispiel nicht möglich, aus einer auf diese Art verschlüsselten Krankheitsbezeichnung (etwa einer Arztbriefdiagnose) direkt die Hauptdiagnose zu erfassen. Da bei der Verschlüsselung anhand einer Nomenklatur im Prinzip keine relevanten Informationen verloren gehen, ist eine Ermittlung der

Hauptdiagnose oder einer anderen Fragestellung allerdings auf Ebene der codierten Dokumente noch immer möglich. Nomenklaturen sind immer unvollständig und müssen regelmäßig den Änderungen des zu Grunde liegenden Fachgebietes angepasst werden. In der Regel sind Nomenklaturen wesentlich umfangreicher als Klassifikationen, die den gleichen Bereich überdecken, und werden deshalb in möglichst disjunkte Kategorien unterteilt, die semantischen Achsen entsprechen (z.B. Topographie, Prozedur, Kosten); dadurch werden sie mehrdimensional (multiaxial). Eine Nomenklatur führt durch die Zusammenfassung von synonymen Begriffen und die Reduktion auf eine Grundgesamtheit zu einer gewissen Standardisierung, enthält also auch bestimmt Elemente einer Klassifikation. Für statistische Auswertungen ist die Indexierung von Dokumenten wenig geeignet, für das Information Retrieval hingegen unentbehrlich. Die mögliche Standardisierung medizinischer Texte anhand einer Nomenklatur, verbunden mit der nachgeschalteten Ermittlung der in ihnen enthaltenen Hauptdiagnose, macht sich der Ansatz dieser Arbeit im Folgenden zueigen.

Weitere Ordnungssysteme, die Eigenschaften von Nomenklaturen und Klassifikationen verknüpfen, sind für bestimmte Einsatzbereiche denkbar [Kiuchi 95]. Welchen grundlegenden Bedingungen sie genügen müssen, wie unter anderem der Disjunktheit der Klassen und der Vollständigkeit in Bezug auf die Grundgesamtheit, die geordnet werden soll, beschreibt [Klar 97]. Aufgrund dieser und anderer Voraussetzungen sind klinische Terminologien immer in gewissem Maße problematisch [Rector 99]; ein perfektes Ordnungssystem für die Medizin kann es nicht geben.

In den folgenden Abschnitten dieses Kapitels werden die Systematische Nomenklatur der Medizin (SNOMED) als Beispiel für eine Nomenklatur sowie die Internationale Klassifikation der Krankheiten (ICD) und den an die Internationale Klassifikation der Prozeduren in der Medizin (ICPM) angelehnten Operationsschlüssel nach § 301 SGB V (OPS-301), nach denen in Deutschland alle Diagnosen bzw. Operationen verschlüsselt werden müssen, als Beispiele für Klassifikationen vorgestellt. Für die automatische Klassifikation von Diagnosenbezeichnungen in ICD bzw. von Prozeduren in ICPM/OPS-301 unter Verwendung von SNOMED ist ein Verständnis dieser Ordnungssysteme grundlegend – Ein weiteres bekanntes Beispiel für Klassifikationen wäre die Internationale TNM-Klassifikation von Stadien maligner Tumore; Beispiele für Nomenklaturen im Bereich der Medizin wären die READ Codes oder der *Unified Medical Language System* (UMLS)-Metathesaurus, die aber den medizinischen Wortschatz in geringerem Umfang als SNOMED abdecken [Campbell 97].

Ausführlichere Beschreibungen der die wichtigsten Ordnungssysteme in der Medizin finden sich in unter anderem in [Zaiß 02, Graubner 95, Cimino 96a]. Wesentliche Informationen zu ICD und ICPM enthält außerdem [GMDS 97].

2.2 Die Nomenklatur SNOMED

Die *Systematized Nomenclature of Medicine* (SNOMED) ist die wichtigste universelle Nomenklatur in der Medizin. Sie erschien erstmals 1975 in den USA in einer Testversion. Grundlage dafür bildete die *Systematized Nomenclature of Pathology* (SNOP), die 1965 erschienen war und bereits damals so angelegt wurde, dass sie für eine automatische Verarbeitung geeignet war. Bereits vier Jahre nach Erscheinen der Testversion gab das *College of American Pathologists* aufgrund der großen Akzeptanz von SNOMED eine revidierte Version (SNOMED II) heraus, auf deren Basis die bisher einzige in Deutschland erschienene Auflage von Friedrich Wingert entwickelt wurde. Dabei mussten neben der Übersetzung auch wesentliche Erweiterungen und Anpassungen an die im deutschsprachigen Raum übliche medizinische Terminologie durchgeführt werden, bevor die Systematische Nomenklatur der Medizin (SNOMED II) 1984 erscheinen konnte [Wingert 84]. – 1993 wurde eine neue englische Version, *The Systematized Nomenclature of Human and Veterinary Medicine* (SNOMED III bzw. SNOMED *international*), veröffentlicht. Dabei wurden außer inhaltlichen Erweiterungen auch strukturelle Änderungen vorgenommen. Allerdings fehlen noch detaillierte Vorschriften für die Reihenfolge und die eindeutige Verbindung der Codes; Hauptproblem des SNOMED ist daher, dass gleiche Ausdrücke oft auf unterschiedliche Art codiert werden können. Diese Problematik führte zu den neueren Entwicklungen von SNOMED *international*, SNOMED RT (*reference terminology*) [Dolin 01, Spackman98], seit Mai 2000 verfügbar, und SNOMED CT (*clinical terms*) [Stearns 01], die Clinical Terms Version 3 (*Read codes*) [O'Neil 95] des National Health Service in Großbritannien mit SNOMED RT kombiniert [Wang 01]. Einen guten Überblick über die Anforderungen an und Ideen für künftige Weiterentwicklungen medizinischer Nomenklaturen wie SNOMED bietet [Cimino 98].

Die SNOMED ist zur Indexierung medizinischer Dokumente und für das Information Retrieval hervorragend geeignet. Dabei wäre eine zusätzliche Indexierung medizinischer Sachverhalte mittels SNOMED die ideale Ergänzung zur gängigen Codierung nach ICD und ICPM, doch die manuelle Benutzung von SNOMED ist aufgrund der komplexen Indizes äußerst aufwendig. Dadurch gewinnt die automatische SNOMED-Verschlüsselung an Bedeutung, wie

sie bei unserem Algorithmus quasi als Nebenprodukt abfällt und sich separat für Fragestellungen des Retrieval nutzen lässt, da sie die Retrievalmöglichkeiten einer Klassifikation wie der ICD deutlich übertrifft [Elkin 01].

SNOMED II

Die SNOMED II ist eine multiaxiale Nomenklatur, die die sieben Achsen Topographie, Morphologie, Funktion, Ätiologie, Prozedur, Krankheit und Beruf umfasst (Tabelle 2.2.1). Diese sind in sich weitgehend hierarchisch aufgebaut und enthalten zu jeder Vorzugsbezeichnung die gängigsten Synonyme; verwandte und veraltete Begriffe sind gesondert gekennzeichnet. Bei der alphanumerischen Notation, die sechs Stellen enthält, steht an erster Stelle der Buchstabe, der die betreffende semantische Achse kennzeichnet, gefolgt von einer fünfstelligen duodezimalen Zahl, wobei „X“ für 10 und „Y“ für 11 steht, also z.B. „TY4220 Bauchnabel“. Zum Teil werden Verweise auf weitere Einträge gegeben, die mit dem Begriff in einer bestimmten Relation stehen. Ein Ausschnitt aus der deutschen Version des SNOMED II findet sich (mit Ergänzungen durch den Algorithmus) in Tabelle 8 im Anhang.

Achse	**Bezeichnung**	**Beschreibung**
T	Topographie	Begriffe zur Beschreibung von Körperteilen, Organen oder Körperregionen
M	Morphologie	Begriffe zur Beschreibung anomaler Topographie oder morphologischer Veränderungen
F	Funktion	Begriffe zur Beschreibung physiologischer und pathophysiologischer Vorgänge
E	Ätiologie (*Etiology*)	Begriffe zur Beschreibung kausaler Agenzien, Chemikalien und Medikamente
P	Prozedur	Begriffe zur Beschreibung therapeutischer, diagnostischer und administrativer Aktivitäten
D	Krankheit (*Disease*)	Begriffe zur Beschreibung komplexer Krankheitsbilder, von Syndromen und anderen komplexen Begriffen
J	Beruf *(job)*	Berufe entsprechend der „Internationalen Standardklassifikation der Berufe“

Tabelle 2.2.1: Die 7 Achsen von SNOMED II

SNOMED III

Wie schon der Titel *The Systematized Nomenclature of Human and Veterinary Medicine* verdeutlicht, wurde SNOMED III (auch als SNOMED *international* bezeichnet) inhaltlich wesentlich erweitert. Der Begriffsumfang hat sich dabei von 44.000 bei SNOMED II über 84.000 bei der deutschen Version auf 156.965 bei der letzten Version 3.5 von 1993 mehr als verdreifacht und bezieht nun unter anderem die Veterinärmedizin, aber auch Pflegediagnosen und -prozeduren mit ein. Um trotz des umfangreichen Volumens weiterhin eine gute Handhabung zu gewährleisten, wurden strukturelle Änderungen der Nomenklatur vorgenommen. Dabei wurden aus den ursprünglich 7 Achsen 11: Die Achse Ätiologie wurde ersetzt durch *"Living Organisms"*, *"Chemicals, Drugs and Biological Products"* und *"Physical Agents, Activities, and Forces"*; neu hinzu kamen *"Social Context"* und *"General Linkage / Modifiers"*. Letztere fasst dabei auch viele Modifikatoren zusammen, die in SNOMED II innerhalb der 7 Achsen bereits teilweise vorhanden waren, aber nur lokale Gültigkeit besaßen. (Beispiele für Modifikatoren wären „rechts" oder „lateral" in Bezug auf Ausdrücke der Topographie, „Verdachtsdiagnose" oder „Ausschluss von" für Diagnosen usw.) Dadurch sowie durch zusätzliche Glossare für einige medizinische Fachbereiche können komplexe medizinische Sachverhalte besser als bisher abgebildet und verknüpft werden, so dass computerbasierte Codierungen von Diagnosen oder Patientenakten mit SNOMED III deren Inhalt sehr getreu widerspiegeln [Rothwell 95]. Die Achsen von SNOMED III sind wie die seiner Vorläuferversion großteils hierarchisch angeordnet; dadurch können implizite semantische Verbindungen für die automatische Codierung genutzt werden [Lussier 98], um die Problematik der unterschiedlichen Codiermöglichkeiten teilweise zu überwinden. Deutlich ausgeweitet wurde die Achse *"Diseases"*, die nun auch die aktuellen Versionen der Diagnosenklassifikationen ICD-O (Onkologie) und ICD-9 berücksichtigt, aus zeitlichen Gründen jedoch nicht mehr die ICD-10. Einzelne medizinische Bereiche sind noch unterrepräsentiert, wie z.B. die Intensivbehandlung. Die Notation von SNOMED III ist hexadezimal (Ziffern 0-9, Buchstaben A-F). SNOMED *"international"* ist für den multilingualen Ausbau gedacht und inzwischen in viele Sprachen übersetzt; eine deutsche Ausgabe existiert derzeit leider nicht.

Achse	Bezeichnung	Beschreibung
T	*Topography*	Begriffe zur Beschreibung von Körperteilen, Organen oder Körperregionen
M	*Morphology*	Begriffe zur Beschreibung anomaler Topographie oder morphologischer Veränderungen
F	*Function*	Begriffe zur Beschreibung physiologischer und pathophysiologischer Vorgänge
A	*Physical Agents, Activities, and Forces*	Physikalische Agenzien, Aktivitäten und Kräfte
C	*Chemicals, Drugs, and Biological Products*	Chemikalien, Medikamente und Biologische Produkte
L	*Living organisms*	Lebende Organismen
P	*Procedure*	Begriffe zur Beschreibung therapeutischer, diagnostischer und administrativer Aktivitäten
D	*Disease*	Begriffe zur Beschreibung komplexer Krankheitsbilder, von Syndromen und anderen komplexen Begriffen
J	*Job*	Berufe entsprechend der „Internationalen Standardklassifikation der Berufe“
S	*Social Context*	Soziales Umfeld
G	*General Linkage / Modifiers*	Allgemeine Modifikatoren, fassen Informationsqualifikatoren und syntaktische Links zusammen

Tabelle 2.2.2: Die 11 Achsen von SNOMED III

2.3 Die internationale Klassifikation von Krankheiten ICD

Die internationale Klassifikation der Krankheiten, Verletzungen und Todesursachen ICD ist die einzige weltweit verwendete Klassifikation in der Medizin und stellt eine allgemein anwendbare Einteilung von Krankheitsbegriffen, Symptomen und Verletzungen dar.

Ihre Wurzeln reichen zurück bis auf den ersten internationalen statistischen Kongress 1853, bei dem die Entwicklung einer solchen Klassifikation beschlossen wurde. Das erste Verzeichnis enthielt 139 Krankheitsbegriffe und Todesursachen; es wurde 1855 verabschiedet und in

der Folge regelmäßig angepasst. – Die erste Version der eigentlichen ICD (ICD-0) stammt aus dem Jahr 1893 und umfasst 3 Einteilungen von Todesursachen mit 44, 99 und 161 Klassen. Seither wurde dieses Werk ungefähr alle 10 Jahre revidiert. Seit 1946 hat die Weltgesundheitsorganisation (WHO) die regelmäßige Revision der ICD übernommen, die seit 1948 (ICD-6) auch Krankheitsbegriffe verzeichnet und in der Folge den Namen Internationale Klassifikation der Krankheiten, Verletzungen und Todesursachen trug. Seit 1975 (ICD-9) wird sie vor allem als Ordnungssystem für Diagnosen zum Erschließen von Krankenakten verwendet. Die neueste Version der ICD (ICD-10) wurde inhaltlich erweitert, was sich in ihrem Namen „Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme“ ausdrückt. Sie trat am 1. 1. 1993 in Kraft. Für die Betreuung der deutschsprachigen Fassung der ICD ist das Deutsche Institut für Medizinische Dokumentation und Information (DIMDI) als Institut des Bundesministers für Gesundheit zuständig, das die deutsche Ausgabe der ICD-10 1994 herausgegeben hat, die seit 1998 für die Diagnosencodierung gesetzlich vorgeschrieben ist.

Da die ICD von der WHO entsprechend internationalen Bedürfnissen gepflegt wird, enthält sie auch Klassen, die in Deutschland völlig unerheblich sind. Andererseits sind Klassen, die bei uns häufig auftretende Krankheiten repräsentieren, die daher differenzierter aufgeschlüsselt werden könnten, zu grob eingeteilt. Aus diesem Grund wurden für einzelne Fachbereiche oder auch klinikumsintern eigene Klassifikationen erstellt, die auf der ICD basierend eine Feineinteilung erlauben. Problematisch ist weiterhin, dass die ICD nicht in allen Fällen die eindeutige Zuordnung einer Diagnose festlegt. Als Hilfsmittel vor allem für Dokumentations-, statistische und künftig Abrechnungszwecke hat die ICD gleichwohl auch klinisch bedeutsame Aspekte, so etwa im frühzeitigen Erkennen des Auftretens von Epidemien [Tsui 01, Espino 01].

Weil in der Testphase des Algorithmus größere Mengen an Testdaten, also von Ärzten über mehrere Jahre und alle Fachrichtungen hinweg codierte Diagnosen, nur in ICD-9 klassifiziert vorlagen und die Evaluation des Systems auf diese Daten zurückgreift, soll im Folgenden zunächst die ICD-9 kurz erläutert werden, bevor die ICD-10 dargestellt wird. Für weiterführende Literatur sei nochmals verwiesen auf [Zaiß 02, Gaus 00, GMDS 97], sowie auf [Gersenovic 95].

ICD-9

Die ICD-9 ist eine monoaxiale Klassifikation, die Krankheitsbezeichnungen in 17 übergeordneten Kapiteln zusammenfasst (Tabelle 2.3.1). Diese sind vorwiegend nach topographischen Aspekten in Gruppen unterteilt, in denen mehrere Klassen zusammengefasst sind. Die Klassen selbst sind zunächst durch 912 dreistellige Schlüssel repräsentiert, darunter 126, die nicht weiter untergliedert werden. Diese sogenannte „Dreistellige Allgemeine Systematik" ist international verbindlich. Die meisten Klassen werden weiter in die „Vierstellige ausführliche Systematik" (5.174 Subkategorien) sowie in für einige Subkategorien optionale fünfstellige Schlüssel unterteilt. Die dritte und vierte Stelle der Notationen sind dabei zur leichteren Lesbarkeit durch einen Punkt getrennt. Drei Zusatzklassifikationen für „Äußere Ursachen und Verletzungen" E, „Morphologie der Neubildungen" M und „Faktoren, die den Gesundheitszustand und die Inanspruchnahme von Einrichtungen des Gesundheitswesens beeinflussen" V ergänzen die Klassifikation. Für bestimmte Krankheiten ist eine Doppelklassifikation sowohl nach ätiologischen als auch nach topographischen Gesichtspunkten möglich, allerdings sind diese beiden Klassen dann gelinkt und die Vorzugsbezeichnung, die für die gesetzlich vorgeschriebene Verschlüsselung von Diagnosen bindend ist, ist gekennzeichnet. Ein Ausschnitt aus der ICD-9 findet sich in Tabelle 10 im Anhang. Für die automatische Codierung steht ein sogenanntes alphabetisches Verzeichnis zur Verfügung, in dem ca. 45.000 Krankheitsbegriffe den korrespondierenden ICD-Klassen zugeordnet sind; ein Ausschnitt daraus ist ebenfalls im Anhang in Tabelle 11 aufgelistet.

Kapitel	**Bezeichnung**	**Dreisteller**
I	Infektiöse und parasitäre Krankheiten	001-139
II	Neubildungen	140-239
III	Endokrinopathien, Ernährungs- und Stoffwechselkrankheiten sowie Störungen im Immunsystem	240-279
IV	Krankheiten des Blutes und der blutbildenden Organe	280-289
V	Psychiatrische Krankheiten	290-319
VI	Krankheiten des Nervensystems und der Sinnesorgane	320-389
VII	Krankheiten des Kreislaufsystems	390-459
VIII	Krankheiten der Atmungsorgane	460-519
IX	Krankheiten der Verdauungsorgane	520-579

X	Krankheiten der Harn- und Geschlechtsorgane	580-629
XI	Komplikationen der Schwangerschaft, bei Entbindung und im Wochenbett	630-676
XII	Krankheiten der Haut und des Unterhautzellgewebes	680-709
XIII	Krankheiten des Skeletts, der Muskeln und des Bindegewebes	710-739
XIV	Kongeniale Anomalien	740-759
XV	Bestimmte Affektionen, die ihren Ursprung in der Perinatalzeit haben	760-669
XVI	Symptome und schlecht bezeichnete Affektionen	780-799
XVII	Verletzungen und Vergiftungen	800-999

Tabelle 2.3.1: Kapitel der ICD-9

ICD-10

Die als ICD-10 bezeichnete Internationale Statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme unterscheidet sich von der ICD-9 im Wesentlichen durch ihren Umfang, da sie, wie es der Name zum Ausdruck bringt, auch weitere Gebiete des medizinischen Umfeldes abdeckt. Sie umfasst nun 21 Kapitel (Tabelle 2.3.2), die in 2.036 dreistellige Kategorien und 12.160 vierstellige Subkategorien untergliedert werden. Um diese Notation auf vier Stellen beschränken zu können, wurde ein alphanumerischer Schlüssel eingeführt, in dem ein Buchstabe die erste Stelle des Codes repräsentiert. Durch die maximal möglich Vergabe von 26.000 Schlüsseln bleibt somit ein großer Spielraum für künftige Erweiterungen. Die Sondersystematiken E und V der ICD-9 sind in die Allgemeine Systematik der ICD-10 integriert; mit der Morphologie von Neubildungen befasst sich eine separate Klassifikation. Das Kreuz-Stern-System, welches eine Doppelklassifikation nach Ätiologie (z.B. Lungenentzündung als Entzündung) und Topographie (z.B. Lungenentzündung als Lungenerkrankung) ermöglicht und damit praktischen Anforderungen Rechnung trägt, wurde gegenüber der 9. Revision erheblich erweitert. Konzeptionell stellt sich die ICD-10 als Mutterklassifikation für andere krankheits- und gesundheitsbezogene Klassifikationen expliziter als frühere ICD-Versionen zur Verfügung. Mit ca. 60.000 ausformulierten Einträgen umfasst das alphabetische Verzeichnis eine Fülle an synonymen Krankheitsbezeichnungen.

Kapitel	Bezeichnung	Dreisteller
I	Bestimmte infektiöse und parasitäre Krankheiten	A00-B99
II	Neubildungen	C00-D48
III	Krankheiten des Blutes und der blutbildenden Organe sowie bestimmte Störungen mit Beteiligung des Immunsystems	D50-D89
IV	Endokrine, Ernährungs- und Stoffwechselkrankheiten	E00-E90
V	Psychische und Verhaltensstörungen	F00-F99
VI	Krankheiten des Nervensystems	G00-G99
VII	Krankheiten des Auges und der Augenanhangsgebilde	H00-H59
VIII	Krankheiten des Ohres und des Warzenfortsatzes	H60-H95
IX	Krankheiten des Kreislaufsystems	I00-I99
X	Krankheiten des Atmungssystems	J00-J99
XI	Krankheiten des Verdauungssystems	K00-K93
XII	Krankheiten der Haut und der Unterhaut	L00-L99
XIII	Krankheiten des Muskel-Skelett-Systems und des Bindegewebes	M00-M99
XIV	Krankheiten des Urogenitalsystems	N00-N99
XV	Schwangerschaft, Geburt und Wochenbett	O00-O99
XVI	Bestimmte Zustände, die ihren Ursprung in der Perinatalperiode haben	P00-P96
XVII	Angeborene Fehlbildungen, Deformitäten und Chromosomenanomalien	Q00-Q99
XVIII	Symptome und abnorme klinische und Laborbefunde, die anderorts nicht klassifiziert sind	R00-R99
XIX	Verletzungen, Vergiftungen und bestimmte andere Folgen äußerer Ursachen	S00-T98
XX	Äußere Ursachen von Morbidität und Mortalität	V01-Y98
XXI	Faktoren, die den Gesundheitszustand beeinflussen und zur Inanspruchnahme des Gesundheitswesens führen	Z00-Z99

Tabelle 2.3.2: Kapitel der ICD-10

2.4 Die internationale Klassifikation von Prozeduren ICPM und der Operationsschlüssel nach § 301 SGB V

Die Internationale Klassifikation von Prozeduren in der Medizin (ICPM) wurde 1978 von der WHO für Forschungs- und Testzwecke herausgegeben. Diese Klassifikation wurde von anderen Ländern teilweise übernommen, oder diente als Basis für die Entwicklung nationaler Klassifikationen. Von den Prozedurklassifikationen, die von der ICPM abgeleitet wurden, hat die ICD-9-CM (*clinical modification*) aus den USA, die sich im Wesentlichen auf Operationen beschränkt, die weiteste Verbreitung gefunden. Weil die notwendige Abstimmung für eine international gültige Klassifikation im Bereich der Verfahren in der Medizin, wo in verschiedenen Ländern sehr unterschiedliche Standards herrschen, äußerst aufwendig ist und dieser Bereich zudem auf Grund des medizinisch-technischen Fortschritts besonders raschen Veränderungen unterliegt, konnte die WHO für die ICPM keinen dem der ICD vergleichbaren Revisionsdienst realisieren. Da außerdem die Bedeutung und das Einsatzgebiet im Vergleich zur ICD nachrangig ist, wurden die Arbeiten an der ICPM eingestellt.

Die Notation der ICPM ist sechsstellig und im Wesentlichen numerisch, lediglich in der 5. und, soweit sie überhaupt besetzt ist, 6. Stelle treten vereinzelt Buchstaben auf. Die ICPM ist wie die ICD monoaxial und hierarchisch strukturiert; dabei spiegeln die erste sowie die ersten 4, 5 oder 6 Stellen des Schlüssels verschiedene hierarchische Ebenen wider (auf Ebene der 2. und 3. Stelle sind zusammengehörende Bereiche angegeben).

Im Auftrag des Bundesministeriums hat das DIMDI 1994 unter Anlehnung an die deutsche Version des ICPM (ICPM-GE) einen amtlichen Operationsschlüssel nach § 301 SGB V (OPS-301) herausgegeben. Er umfasst sämtliche Klassen der ICPM, die zur Klassifikation von Operationen und anderer für gesetzliche Belange relevanter Prozeduren notwendig sind. Dabei wurde auf eine Kompatibilität bis zur vierten Stelle geachtet, soweit dies auf Grund des Fortschrittes in der Medizin zu erreichen war. In Deutschland müssen seither alle Operationen danach klassifiziert werden. Zur Umsetzung der gesetzlichen Verordnung zu Fallpauschalen und Sonderentgelten wurde der OPS-301 im Jahre 1995 angepasst, und zur Einführung des DRG-Systems nochmals überarbeitet und wesentlich erweitert. Die seit dem 1. Januar 2002 in der stationären Krankenhausversorgung zu verwendende Version 2.1 ist sechsstellig und dekadisch. Sie ist unterteilt in fünf Kapitel (Tabelle 2.4; wegen der geforderten Kompatibilität zur ICPM sind nicht alle Stellen besetzt). Ein amtliches alphabetisches Verzeichnis wie für die ICD liegt für die aktuelle Version des OPS-301 noch nicht vor, ist aber in Vorbereitung.

Das DIMDI verfolgt aktuelle Entwicklungen im Bereich der Prozedurenklassifikation in anderen Ländern, um langfristig eine Nachfolgeklassifikation für den OPS-301 zu finden. Die französische *Classification commune des actes médicaux* (CCAM) und das amerikanische *Procedure Coding System* (PCS) sind mögliche Kandidaten.

Kapitel	**Bezeichnung**
1	Diagnostische Maßnahmen
3	Bildgebende Diagnostik
5	Operationen
8	Nichtoperative therapeutische Maßnahmen
9	Ergänzende Maßnahmen

Tabelle 2.4: Kapitel der OPS-301 (Version 2.1)

3 Methodik des Indexier-Algorithmus

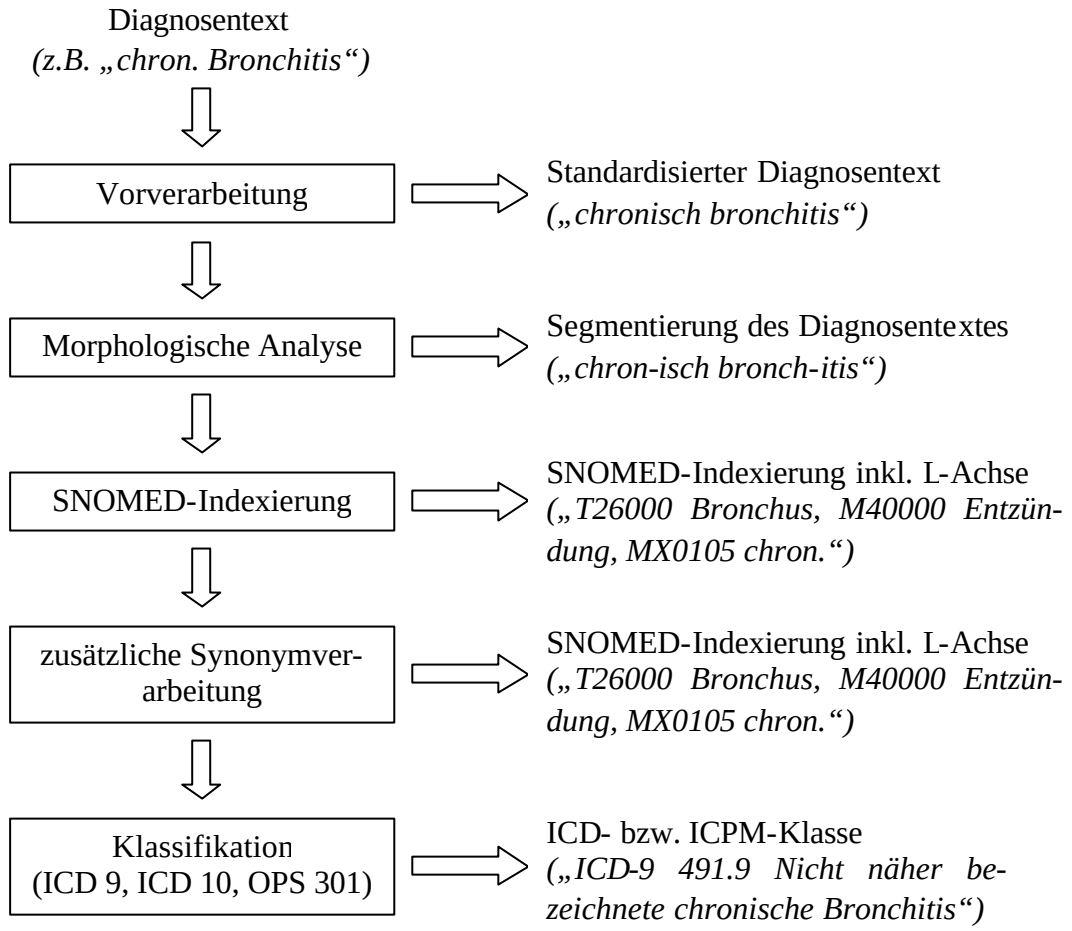
Der aktuelle Stand der automatischen Indexierung und Klassifikation von Krankheitsbezeichnern wurde in Kapitel 1.3 in Grundzügen dargestellt. Entsprechend der dortigen Aufteilung ist auch das im Rahmen dieser Arbeit entwickelte System, das als MedSearch bezeichnet wird, aufgeteilt in einen Algorithmus zur **Indexierung** von Diagnosen mittels SNOMED-Termen (Kapitel 3) und ein neuartiges Konzept zum **Retrieval** von Suchanfragen (Kapitel 4) in der so codierten Dokumentenkollektion (im Folgenden ICD-9, ICD-10 oder OPS-301).

Der Indexier-Algorithmus ist in seinen Grundzügen eine Weiterentwicklung des an der Universität Heidelberg entworfenen lexikabasierten Indexierverfahrens LBI, das dort in dem System SALBIDH umgesetzt und evaluiert wurde [Brigl 94, Brigl 95]. In einem Überblick über Verfahren des *Natural Language Processing* (NLP) schneidet dieses System mit am besten ab; für Weiterentwicklungen werden syntaktisch motivierte Ansätze mit lokaler bereichsspezifischer semantischer Analyse empfohlen [Spyns 96].

3.1 Überblick über den MedSearch-Algorithmus

Die Indexierung von Diagnosentexten, d.h. ihre Abbildung (Codierung) in Texte des kontrollierten Vokabulars von SNOMED II, verläuft in mehreren Teilschritten, die sich linear folgen. Dabei sind auf jeder Teilstufe Zwischenergebnisse abrufbar (etwa die morphologische Analyse des Textes; zum technischen Aufbau und den Schnittstellen des Systems siehe Kapitel 5.1) und können für andere Applikationen verwendet werden. An die SNOMED-Indexierung schließt sich über den Zusatzschritt einer weiteren Synonymverarbeitung (Abschnitt 3.5) die ICD-Klassifikation des Textes mittels des Retrievalverfahrens an, das in Kapitel 4 vorgestellt wird und den MedSearch-Algorithmus beendet.

Zum besseren Überblick hier eine graphische Verdeutlichung der Abfolge der einzelnen Schritte (Skizze 3.1), sowie eine kurze Erläuterung:



Skizze 3.1: Schritte zur SNOMED-Indexierung und ICD- bzw. OPS 301-Klassifikation. Auf jeder Ebene lassen sich Teilergebnisse abrufen.

Vorverarbeitung. Bei der Vorverarbeitung der Diagnosentexte findet eine Wortstandardisierung auf syntaktischer Ebene statt; des weiteren werden unter anderem Abkürzungen aufgelöst. Die Vorverarbeitung besteht aus vielen kleinen Einzelschritten, die in Abschnitt 3.2 detailliert aufgeführt werden.

Morphologische Analyse: Die morphologische Analyse (Abschnitt 3.3) dient dazu, mit Mitteln der Lemmatisierung und Dekomposition die Einzelworte des Diagnosentextes in morphologische Grundformen (Morpheme) aufzutrennen (segmentieren, parsen). Bei dieser Segmentierung bleiben ursprüngliche Zusammenhänge von Morphemen als Wörter sowie die Reihenfolge dieser Wörter im Diagnosentext erhalten. Zu Grunde liegen dem Teilalgorithmus ein umfangreiches Morphemlexikon, das zum kleineren, manuell erstellten Teil den Grundwortschatz der Medizin abdeckt, sowie ein einfaches Wortmodell, das grammatikalisch verschiedene Arten von Morphemen kennt und anhand dessen Wörter zerlegt werden können.

SNOMED-Indexierung: Im nächsten Schritt findet auf Basis der morphologischen Dekomposition des Diagnosentextes eine Indexierung mittels des kontrollierten Vokabulars von SNOMED II statt (Abschnitt 3.4). Da viele Begriffe, die aus Randbereichen der Medizin stammen oder Spezialwissen erfordern, darin nicht verzeichnet sind, wurde den 7 Achsen des SNOMED II eine zusätzliche umfangreiche Lexemachse (Notation L + 5stellige dekadische Zahl) hinzugefügt, in der sämtliche bei der Wortdekomposition der Dokumentenkollektion nicht wiederzufindenden eigenständig bedeutungstragenden Bezugseinheiten (Lexeme) verzeichnet sind. Zur Indexierung muss eine morphologische Analyse des gesamten SNOMED II vorliegen, die vorab erfolgt.

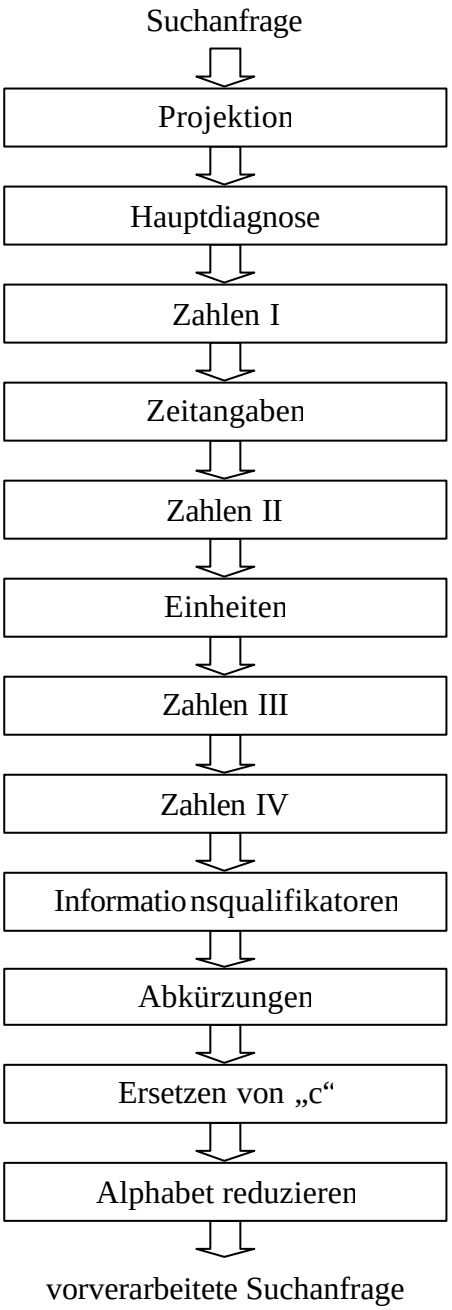
Zusätzliche Synonymverarbeitung: Durch die große Zahl an gleichbedeutenden Bezeichnungen, die SNOMED II unter den einzelnen Codes erfasst, wird bereits eine sehr weitgehende Synonymbehandlung erreicht. Eine weitere Gleichsetzung von Begriffen ist in Einzelfällen sinnvoll (Abschnitt 3.5). Allerdings ist es schwierig abzuwägen, in wiefern das Projizieren unterschiedlicher Termini auf eine Vorzugsbezeichnung in weiterem Umfang wirklich sinnvoll ist. Verschiedene Fehlerquellen stellen sich einer zu weitgehenden Synonymisierung entgegen, so dass momentan die zusätzliche Synonymverarbeitung sehr restriktiv gehandhabt wird.

Klassifikation (Kapitel 4): Der eigentliche Klassifikationsalgorithmus beruht auf der vorhergehenden Indexierung der Suchanfrage mittels SNOMED II. Er nützt das der Architektur von SNOMED II zu Grunde liegende implizite Wissen über Hierarchien medizinischer Begriffe und die weitgehend semantische Disjunktheit der 7 Achsen in einer heuristischen Weise, um die Retrievalqualität durch Hintergrundwissen zu erhöhen. Dabei stehen für ICD-9, ICD-10 und OPS-301 Lexika mit zusätzlichen Beispieldiagnosen bzw. -prozeduren zur Verfügung, die mit dem gleichen Indexieralgorithmus verschlüsselt in Form von SNOMED-Codes vorliegen.

Ein Schema für einen Algorithmus, der mit dem hier skizzierten in wesentlichen Teilen übereinstimmt (Entfernen von Stopwörtern, Dekomposition und Reduktion auf Grundformen, Gruppieren von Wörtern zu Konzepten, morphosemantische Abstandsdefinition), darüber hinaus aber noch weitergehende Vorschläge für Indexierung und Retrieval bringt, die für unseren Ansatz nicht relevant waren, wird in [Baud 98a] vorgestellt.

3.2 Vorverarbeitung der Suchanfrage

Bei der Vorverarbeitung werden Suchanfragen und Dokumente syntaktisch aufbereitet, um bereits auf unterster Ebene eine gewisse Standardisierung von Texten zu erreichen, bei der für die Klassifikation überflüssige Information verworfen und relevante Information (z.B. Abkürzungen) herausgearbeitet wird. Sie besteht aus einer Reihe einzelner Arbeitsschritte, die wie in Skizze 3.2 dargestellt linear aufeinander folgen und im Weiteren erläutert werden sollen.



Skizze 3.2: Vorverarbeitungsschritte

Projektion

Der erste Teilabschnitt der Vorverarbeitung projiziert den zu analysierenden Text, eine Kette aus ASCII-Zeichen, auf einen reduzierten Satz an Buchstaben, Zahlen und Sonderzeichen. Dabei werden unter anderem Akzente eliminiert (Beispiel: „é“ wird zu „e“). Viele Sonderzeichen werden auf das Leerzeichen abgebildet ($, @, +, #, Zeilenumbruch usw.) und schaffen somit eine Abgrenzung von Zeichenketten. Gesondert ersetzt werden „°“ durch „Grad“ und „&“ durch „und“. Nach diesem Schritt verbleiben die Zeichen a…z, A…Z, ä ö ü ß Ä Ö Ü, 0…9, %, µ, !, ?, „, ;, :, ., -, ', ". – Da in der morphologischen Analyse im Wesentlichen nur noch Wörter zerlegt werden, die aus a…z bestehen, werden die anderen Zeichen (bis auf das Leerzeichen) in der weiteren Vorverarbeitung schrittweise analysiert und entfernt. Eine detaillierte Definition des Projektionsmechanismus kann der zu Grunde liegenden Ersetzungstabelle **(1) Alphabet** entnommen werden (Ausschnitte aus den Tabellen siehe Anhang).

Hauptdiagnose

Viele Kurzdiagnosen enthalten außer der Hauptdiagnose die Angabe einer weiteren Ursache, Komplikation oder eines Nebenbefundes. Einiger dieser Konstrukte sind rein syntaktisch erkennbar. Da es für die Indexierung nicht relevant ist, ob eine oder mehrere Diagnosen in der Suchanfrage vorhanden sind, wohl aber für die Klassifikation, kann durch ein Flag angegeben werden, ob Nebendiagnosen entfernt oder belassen werden sollen. Prinzipiell wäre es sinnvoll, in einem solchen Fall alle vorkommenden Diagnosen getrennt zu klassifizieren und dem Benutzer zur Auswahl vorzulegen. Für die vollautomatische Klassifikation, bei der Rückfragen nicht möglich sind (dies kann zum Beispiel bei der Auswertung einer großen Anzahl von Diagnosen zu statistischen Zwecken der Fall sein), wäre neben syntaktischen Unterscheidungskriterien eine Beurteilung anhand von Charakteristika wie der medizinischen Relevanz einer bestimmten Diagnose oder ihrer Häufigkeit möglich, für die zusätzliche Datensätze erforderlich wären.

Folgende Konstrukte werden derzeit in dieser Reihenfolge behandelt:

- Alles, was in Klammern steht, wird entfernt.
- Diagnose „A mit B“ wird zu „A“.
- Diagnose „A wegen B“ wird zu „B“.

Formulierungen wie „A bei B“ und ähnliche Ausdrücke legen noch nicht fest, ob „A“ oder „B“ die zu klassifizierende Diagnose ist. Bisher wird daher beides gemeinsam weiter verarbeitet. Eine Verfeinerung dieses Ansatzes wäre durchaus denkbar.

Behandlung von Zahlen I

Die Behandlung von Zahlen ist in vier Teilschritten gegliedert, in deren erstem Zeichenketten, die nur aus 0…9 bestehen (Ganzzahlen) erkannt und in eckigen Klammern zusammengefasst werden. So wird beispielsweise „2. 7. 1999“ zu „[2]. [7]. [1999]“ und „1.5 mg“ zu „[1].[5] mg“. Die unterschiedliche Bedeutung, die Interpunktionen in diesem Zusammenhang haben können (so kann ein Punkt bei Datumsangaben stehen oder als Dezimaltrennzeichen), macht hier eine Unterteilung der Verarbeitung von Zahlen notwendig.

Zeitangaben

Im nächsten Vorverarbeitungsschritt werden Datumsangaben erkannt und durch den Term „{Zeit}“ ersetzt. Darunter fallen folgende Formulierungen: Tag+Monat+Jahr; Monat+Jahr; Jahr (vierstellig). Tage werden dabei durch die Zahlen 1-31 repräsentiert, Monate durch Zahlen 1-12 oder Einträge aus der Tabelle **(2) Months** (alle Tabellen siehe Anhang), Jahre durch die Ziffernfolgen 00-99 und 1900-2100. Des weiteren werden folgende Angaben ermittelt und durch „{Zeit}“ ersetzt: „{Zeit}...{Zeit}“, „{Zeit}-{Zeit}“, „von {Zeit}“, „vom {Zeit}“, „am {Zeit}“, „seit {Zeit}“, „zum {Zeit}“, „bis {Zeit}“ (jeweils auch mit Großbuchstaben am Anfang). Nicht erkannt werden Uhrzeiten.

Da sich die Relevanz von Datumsangaben in bezug auf die Klassifikation von Diagnosen mittels ICD auf Unterscheidungen wie „akuter Herzinfarkt“ und „alter Herzinfarkt“ beschränkt, werden sämtliche Zeitangaben innerhalb einer Diagnose derzeit als Kennzeichen eines nicht akuten Vorganges gewertet und durch den Begriff „alt“ ersetzt.

An dieser Stelle könnte sich eine Routine zur Erfassung von Aufzählungen anschließen, um etwa Auflistungen mehrerer Diagnosen als solche zu erkennen und diese einzeln weiterzube-

handeln. Diese Idee wurde im Konzept der Protokollierung aller vom Algorithmus vorgenommenen Ersetzungen in einem Pseudocode (technische Details in Kapitel 5.2) bereits insofern aufgegriffen, als sich eine Aufspaltung von Einzeldiagnosen in mehrere Varianten durch eine Anweisung beschreiben lässt. Um die Komplexität des Algorithmus durch das Verlassen des derzeitigen straight-forward-Ansatzes aber nicht zu vervielfachen, wurde ein solcher möglicher Teilschritt bisher nicht verwirklicht und sei nur der Vollständigkeit halber erwähnt.

Behandlung von Zahlen II

Nachdem im vorigen Schritt Datumsangaben erkannt wurden, kann nun ein Punkt oder Komma in Zusammenhang mit einer Zahl als Dezimaltrennzeichen interpretiert werden. Daher werden im zweiten Schritt der Verarbeitung von Zahlen Kommazahlen zusammengefasst, so dass beispielsweise „[1].[5] cm“ zu „[1.5] cm“ wird. Es schließt sich die Behandlung von Einheiten an.

Einheiten

Da in der Medizin für die gleiche physikalische Größe zum Teil unterschiedliche oder spezifische Einheiten verwendet werden, wie etwa Fahrenheit oder Celsius statt der Grundeinheit Kelvin, und Einheiten generell durch Anhänge wie „centi“, „kilo“ usw. um Zehnerpotenzen variieren können, ist eine Normierung notwendig. Wann immer auf eine Zahlenangabe daher eine Einheit folgt, wird diese vom Algorithmus in eine festgelegte und vergleichbare Grundeinheit umgerechnet. So wird etwa „[1.5] cm“ zu „[0.015 m]“, genauso wie „[15] mm“. Die einzelnen Faktoren und Summanden, die zu der Ersetzung notwendig sind, können der Tabelle **(3) Units** entnommen werden. Im Anschluss wird das Sonderzeichen „µ“, so noch vorhanden, durch ein Leerzeichen ersetzt. – Eine große Bedeutung innerhalb der Klassifikation von Diagnosen und Prozeduren kommt diesem Teilschritt sicherlich nicht zu, bei der Verschlüsselung in andere medizinische Ordnungssysteme wie etwa die TNM-Klassifikation ist er hingegen äußerst relevant.

Behandlung von Zahlen III

Weit verbreitet ist im medizinischen Wortschatz die Verwendung von römischen Ziffern. Diese werden im Vorverarbeitungsschritt Zahlen III gesondert behandelt und soweit möglich in arabische Zahlen umgewandelt. Ein typisches Beispiel wäre die Konvertierung von „Chromosom XXIII“ zu „Chromosom [23]“ (eine komplette Liste der zu ersetzenden Ziffern ist in Tabelle **(4) Numbers** enthalten.). Problematisch sind hier allerdings die römischen Zahlenangaben, die durch einen einzelnen lateinischen Buchstaben repräsentiert werden; in der Medizin gebräuchlich sind hier I, V und X. Alle drei können für Abkürzungen („V. Cava“) oder Zahlen („V. Hirnnerv“, „Adipositas Grad I“) stehen und teilweise in feststehenden Begriffen verwendet werden („X-Beine“), so dass es zu Ambiguitäten kommt. Die Entscheidung, welche Interpretation korrekt ist, kann nur im Kontext erfolgen. Bisher werden diese Buchstaben daher bei der (rein syntaktischen) Vorverarbeitung nicht ersetzt; da die für die Klassifikation wichtigsten Verwendungen dieser Einzelbuchstaben in SNOMED, dem Synonymlexikon und der Dokumentenkollektion ICD-verschlüsselter Diagnosen bzw. OPS-301-verschlüsselter Prozeduren festgehalten sind, erfolgt ihre Deutung dann an späterer Stelle.

Behandlung von Zahlen IV

Um die Vorverarbeitung von Zahlen abzuschließen, werden in einem letzten Schritt alle Zahlenangaben auf zweistellige Genauigkeit gerundet, d.h. auf eine Zahl zwischen 10 und 99, multipliziert mit einer ganzzahligen Potenz von 10. So wird beispielsweise „[137 mmHg]“ zu „[140 mmHg]“. Zahlen, die nahe genug beieinander liegen, werden so identifiziert. Bei Werten, bei denen ein Unterschied in der dritten oder weiteren Stellen noch relevant ist (häufig z.B. bei Laborparametern), können sich hier Probleme ergeben. Für die Klassifikation von Diagnosen und Prozeduren haben solche Parameter in der Praxis aber kaum Bedeutung und der Vorteil der Identifikation vergleichbarer Maßzahlen überwiegt.

Informationsqualifikatoren

Vor der folgenden Auflösung von Abkürzungen, spätestens aber vor der morphologischen Analyse ist es sinnvoll, in einem Zwischenschritt medizinisch gebräuchlichen Informationsqualifikatoren wie „V. a.“ (Verdacht auf) zu identifizieren und markieren. Diese auch als Mo-

difikatoren bezeichneten Begriffe sind einer kurzen Liste entnommen, die in gleicher Form der Arbeit von [Brigl 94, Brigl 95] zur Verfügung stand und die in Tabelle **(5) Qualifiers** verzeichnet ist. Sie haben eine semantische Bedeutung in Bezug auf den Kontext des zu klassifizierenden Krankheitsbegriffs, indem sie ihn z.B. als frühere oder Differenzialdiagnose kennzeichnen oder negieren.

Abkürzungen

Da sowohl Kurzdiagnosen des Arztes auf Station als auch längere Arztbriefdiagnosen sehr häufig und zum Teil sehr komplexe Abkürzungen enthalten, ist die ihre korrekte Auflösung eine der wichtigsten Aufgaben in der Verarbeitung solcher Texte. Dabei sind zwei Arten von Abkürzungen zu unterscheiden: Abkürzungen mit einem Punkt am Ende, die in der Regel ein längeres Wort trunkieren und in Kurzdiagnosen zahlreich sind, und Akronyme (ein Akronym ist ein aus den Anfangsbuchstaben mehrerer Wörter gebildetes Wort).

Ein gut bekanntes Problem insbesondere im Falle von Akronymen ist die Mehrdeutigkeit medizinischer Abkürzungen [Wren 02]. Als Standardbeispiel für beliebig viele andere Fälle sei das Kürzel „HWI“ genannt, das sowohl für „Hinterwandinfarkt“ als auch für „Harnwegsinfekt“ stehen kann. Je nach Fachabteilung (so bekannt) ist es daher sinnvoll, bestimmte Varianten zu bevorzugen; dabei muss bei mehrdeutigen Abkürzungen die Vorzugsbezeichnung für die einzelnen Fachrichtungen gekennzeichnet sein. Eine weitere Möglichkeit wäre auch hier, beide Varianten getrennt weiterzuverfolgen und dem Benutzer (bei der semiautomatischen Verschlüsselung) mehrere Ergebnisse anzubieten, eine Variante die der Algorithmus indes nicht verfolgt. Statt dessen wird im Zweifelsfall eine der Möglichkeiten vorgezogen.

Andere Probleme stellen sich bei der Trunkierung von Wörtern (Abkürzungen auf Punkt). Überschneidungen mit dem Punkt als Satzzeichen sind möglich, führen jedoch selten zu Mehrdeutigkeiten. Schwieriger zu behandeln sind Abkürzungen, die sich über mehrere Worte in Folge erstrecken (beispielsweise „i. v.“ für „intravenös“) oder nur im Kontext mit weiteren Worten ihren Sinn erschließen lassen (etwa „E. Coli“ für „Escherichia Coli“). Ein großes Problem sind vor allem auch „wilde“ Abkürzungen, die vom Arzt willkürlich und nach eigenem Erachten generiert werden; sie finden sich in keiner Datenbank wieder, es sei denn in einem automatisch dazulernenden System (hierzu wäre jedoch ein Feedback des Benutzers notwendig, welches das System MedSearch nicht voraussetzt). Dennoch kann der Algorithmus sol-

che Abkürzungen teilweise auflösen. Da die Muster, nach denen abgekürzt wird, sich in der Regel ähneln, entstand die Idee, nicht nur ganze Wörter wie „Schlafkrkht." zu betrachten, sondern auch die Wortenden; so lässt sich das Wort „Schlafkrankheit" ermitteln, indem am Wortende „-krkht." zu „-krankheit" expandiert wird. Allerdings hat auch dieser Ansatz Grenzen. Kurze und unspezifische Wortenden wie beispielsweise „-ol." können oft ohne den Kontext nicht mehr eindeutig zugeordnet werden („-olisch"?, „-ologisch"?). Daher wird dieser Abkürzungstyp etwas restriktiv genutzt.

Tröstlich ist, dass in vielen Fällen von „wilden" Abkürzungen mit einem Punkt am Ende nur trunkiert wurde und der Wortstamm mit der wesentlichen medizinischen Information ganz oder teilweise erhalten und der weiteren morphologischen Verarbeitung zugänglich bleibt. So wird bei unauflösbaren Abkürzungen die Zeichenkette ohne den folgenden Punkt einfach beibehalten. Bei unbekannten vermeintlichen Akronymen handelt es sich zum Teil lediglich um Großschreibweisen z.B. von Eigennamen, so dass auch sie für die weiteren Schritte bewahrt werden.

Die verwendete Tabelle **(6) Abbreviations** enthält derzeit 710 Abkürzungen auf Punkt (inklusive Wortenden) und Akronyme. Zahlreiche zusätzliche Akronyme sind im weiteren Lerndatensatz (SNOMED und den ICD- bzw. OPS-301-Einträgen, vgl. Abschnitt 5.3) enthalten, werden aber nicht explizit aufgelöst, da sie längst den Charakter eines Eigennamens angenommen haben. Ein ganz typisches Beispiel hierfür wäre das Akronym „AIDS".

Ersetzen von „c"

Viele medizinische Termini, die ein c enthalten, besitzen eine deutsche und eine lateinische Schreibweise. Um Varianten wie „Karzinom" und „Carcinom" zu vereinheitlichen, werden in aller Regel z und k auf c abgebildet. Eine solche Vereinfachung löst zwar dieses Problem, führt aber in anderen Fällen zur Identifikation sehr unterschiedlicher Begriffe. So ist etwa der Begriff „Zoster" nach Abbildung auf „Coster" von dem lateinischen Wort für Rippe „Costa" nur noch durch die im Allgemeinen unerheblichen Endungen -er bzw. -a zu trennen.

MedSearch verfolgt hier einen anderen, phonetischen Ansatz, indem in umgekehrter Weise kontextabhängig c durch z bzw. k ersetzt wird. Vor den heller klingenden Vokalen e, i, ä, ö und ü wird ein lateinisches c in aller Regel als z gesprochen, vor a, o, u und Konsonanten hin-

gegen als k. Da sich die deutsche (oder verdeutschte) Schreibweise dieser Begriffe an diese Ausspracheregeln hält, ist die exaktere Methode, nicht z und k durch Ersetzung zu eliminieren, sondern umgekehrt den Buchstaben c. Lediglich als ch, ck oder am Ende eines Morphems im Morphemlexikon bleibt er bestehen (beim letzten Punkt, da verschiedene Fortsetzungen möglich sind). Zweideutigkeiten kommen so nur noch äußerst selten vor.

Typische Beispiele, die diesen Ersetzungsmechanismus verdeutlichen können, sind etwa „Calcium“, das zu „Kalzium“ wird, „Caecum“ zu „Zaekum“, „Costa“ zu „Kosta“, „Ulcus“ zu „Ulkus“ aber „Ulcera“ zu „Ulzera“ usw. Im Morphemlexikon muss sich im letzten Fall der Wortstamm „Ulc“ befinden, der sowohl für „Ulk“(-us) als auch für „Ulz“(-era) steht.

Projektion auf ein reduziertes Alphabet

Im letzten Teilschritt der Vorverarbeitung kann die Anzahl an Zeichen, die der weiteren Verarbeitung zur Verfügung stehen sollen, nochmals reduziert werden. Insbesondere ist für die morphologische Zerlegung von Wörtern deren Groß- oder Kleinschreibung nicht von Belang. (Allerdings gibt es durchaus Wörter, die groß- oder kleingeschrieben eine unterschiedliche Bedeutung haben, aber diese Fälle sind selten.) Außerdem werden die Umlaute ä, ö, ü und ß zu ae, oe, ue und ss; damit jedoch bei der morphologischen Zerlegung keine Aufteilung in a-e, s-s usw. vorgenommen werden kann, bleiben diese Buchstaben zunächst eingeklammert: [ae], [ss]. Zahlangaben bleiben erhalten, sie werden auch morphologisch nicht zerlegt, wurden aber in der Vorverarbeitung bereits ausführlich behandelt und stehen der SNOMED-Indexierung direkt zur Verfügung.

Eine besondere Bedeutung haben die Zeichen „-“ am Anfang und „*“ am Ende einer Zeichenkette, da sie in SNOMED II dafür stehen, dass der entsprechende Eintrag lediglich einen Präfix wie z.B. „prae*“ oder einen Suffix wie z.B. „-itis“ darstellt, nicht ein komplettes Wort. Da die morphologische Zerlegung aber Worte voraussetzt, die einen Wortstamm besitzen, wird hier ein imaginärer, in sich unzerlegbarer Wortstamm angefügt. Nach der Zerlegung des auf diese Art künstlich generierten Wortes wird dieser Wortstamm später wieder eliminiert.

Typisch ist weiterhin die Verwendung des Zeichens „~“, um Worte abzukürzen, deren Endungen sich wiederholen. Diese Zeitersparnis bei der Eingabe führt zu Problemen bei der automatischen Verarbeitung, da für das System meist nicht erkennbar ist, für welche Auslassung der

Bindestrich steht. So wird er derzeit an dieser Stelle des Algorithmus entfernt. Zurück bleibt ein Teil des Wortes und damit ein Teil der Information, der morphologisch zerlegt werden kann, wenn es sich um einen Wortstamm handelt („Leber- und Lungenfiliae“ wird zu „Leber und Lungenfiliae“, „Mitral- und Aortenvitium“ zu „Mitral und Aortenvitium“), oder verworfen wird, wenn es sich um eine Vorsilbe handelt („Ober- und Unterschenkelfraktur“ wird zu „Unterschenkelfraktur“, „In- und expiratorischer Stridor“ zu „Expiratorischer Stridor“). Die Bedeutungsverschiebung ist in jedem der Beispiele beachtlich. Charakteristisch für solche Konstrukte ist das Bindewort „und“ nach der Auslassung, die in einem klinischen Testdatensatz von 10.000 Entlassungsdiagnosen in 1,7 % der Fälle vorkam, was eine Behandlung sehr wünschenswert erscheinen lässt. Das Anfügen eines hypothetischen Wortstammes, der nach der morphologischen Zerlegung entfernt wird, würde hier zumindest das vorzeitige Entfernen von Vorsilben verhindern, aber im Prinzip sollte eine eingehendere Analyse der auf die Auslassung folgenden Worte diese zu ermitteln versuchen. –

Eine weitergehende Projektion von Zeichen, wie etwa f ⇨ ph, y ⇨ ie oder ähnliches findet nicht statt. Der damit zu gewinnenden Vereinheitlichung vieler Wörter, die in den meisten Fällen ohnehin nur in einer der Schreibweisen existieren, steht ein inadäquater oder zumindest nicht unerheblicher Informationsverlust entgegen. Die Projektion von c auf k und z sowie die Sonderbehandlung von Umlauten durch Klammerung indes erhöhen die Komplexität einer Implementierung der nun folgenden morphologischen Analyse, so dass ein Verzicht auf diese Sonderregelungen eher plausibel scheinen könnte. –

Generell ist zu betonen dass, auch wenn die vielen Einzelschritte und Details der Vorverarbeitung, die die letzten Seiten schildern, etwas kleinlich erscheinen mögen, die Qualität eines Indexier- und Retrievalalgorithmus sehr viel mehr von einer gründlichen Vorverarbeitung der Suchanfragen und Dokumente abhängt, als dies den Anschein haben mag, und dass hier zum Teil größere Verbesserungen zu erzielen sind als etwa mit einer verfeinerten morphologischen Analyse. Dennoch wird auch im folgenden Abschnitt versucht, den weiteren Ablauf möglichst zu optimieren.

3.3 Morphologische Analyse der vorverarbeiteten Suchanfrage

Worte sind zusammengesetzt aus kleineren sprachlichen Einheiten, den *Morphemen* (z. B. Poly-neur-o-path-ie). Dabei können Präfixe (Poly-), Wortstämme (neur, path), Fugenmorpheme (-o-) und Suffixe (-ie) unterschieden werden. Das Regelwerk an Konstruktionsmechanismen von Wörtern aus solchen sprachlichen Grundeinheiten ist die *Morphologie.* Die wichtigsten dieser Wortbildungsprinzipien sind Komposition (Zusammensetzung von Wortstämmen oder Wörtern, ggf. mit Interposition eines Fugenmorphems), Präfiguierung (Modifikation eines Wortstammes oder Wortes durch Voranstellen eines Präfixes), Derivation und Flexion (Modifikation durch Anhängen eines Suffixes). Die *morphologische Analyse* beschreitet den umgekehrten Weg, indem sie Buchstabenketten in Einklang mit den beschriebenen Wortbildungsprinzipien in mögliche Sequenzen von Morphemen zerlegt und damit den ersten Schritt von der Zeichenfolge hin zu Bedeutung und Semantik geht.

Insbesondere durch die enorme Häufigkeit von Komposita in der medizinischen Fachsprache kommt der morphologischen Analyse eine sehr große Bedeutung zu. Ein Lexikon aller möglichen Komposita und Flexionen eines Wortstammes wäre nicht nur von überdimensionalem Umfang, es könnte auch niemals vollständig sein, da durch die stete Weiterentwicklung der Terminologie in der Medizin sowie durch oft willkürliche Wortneuschöpfungen, die nur durch Dekomposition verständlich sind, eine klare Grenzziehung gar nicht möglich ist. Über die Zerlegung in Morpheme treten zudem semantische Verbindungen zwischen Wörtern zu Tage, die zur Definition eines Abstandes für Retrievalfunktionen genutzt werden können. Darum wird im Folgenden ein Algorithmus zur Wortsegmentierung vorgestellt, bei dem die ursprüngliche Reihenfolge von Morphemen innerhalb eines Wortes sowie die Reihenfolge der Wörter im zu analysierenden Dokument erhalten bleiben, da beides einen relevanten Informationsgehalt birgt, der im weiteren Verlauf Verwertung findet.

Die Grundform des Algorithmus wird dabei zusätzlich durch Ausnahmeregelungen für verschiedene spezielle Probleme der morphologischen Analyse modifiziert. Dazu zählen unter anderem Sonderregeln auf Buchstabenebene (Untrennbarkeit von [ae], [ss] usw., Entwicklung von c zu z oder k am Ende von Morphemen, Abbildung von Umlauten auf den Grundvokal um z.B. „Läuse“ in „Laus“+ „-e“ zerlegen zu können); die Umsetzung dieser eher technischen Details wird in Kapitel 5.4 näher erörtert. Das Problem von Idiomen wie „Darmbein“, deren Zerlegung ihre medizinische Bedeutung entstellen würde, wird dadurch umgangen, dass auch nicht atomare Stammformen als Morpheme ins Lexikon aufgenommen werden, mit der Kon-

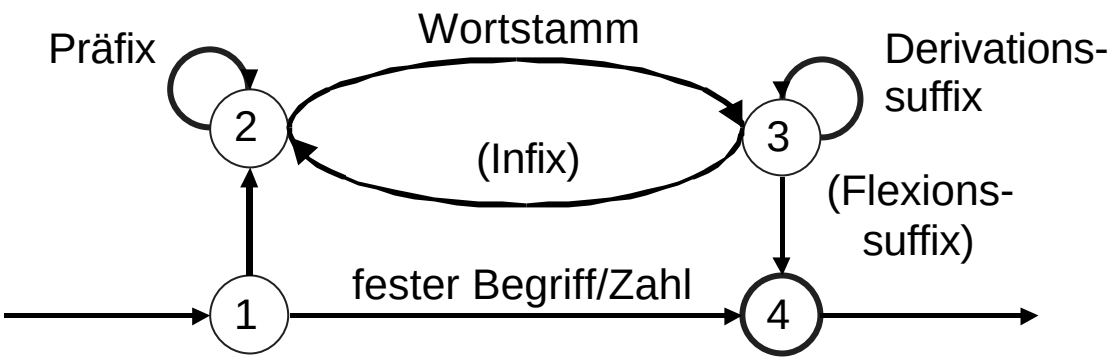
sequenz dass in semantischer Relation stehende Begriffe wie „Blinddarm“ und „Darm“ auf morphologischer Ebene disjunkt bleiben. Diese Problematik könnte nur durch die zusätzliche Speicherung von Relationen (das ganze wäre dann ein *Thesaurus*), die für den Algorithmus allerdings ein wesentlich komplexeres Regelwerk erforderlich machen würden, aufgelöst werden. Zu unspezifische Wortstämme („Ei“, das dem griechischen „Oo“ wie in „Oo-zyt“ entspricht), feststehende, nicht mit anderen Morphemen kombinierbare Begriffe („Basedow“) sowie Zahlen werden im Morphemlexikon als nicht kombinierbar gekennzeichnet und vom Algorithmus nur als Ganzes, nicht aber als Teil eines Wortes analysiert.

Da die morphologische Dekomposition von Wörtern bei der Aufnahme auch nicht atomarer Stammformen in das Morphemlexikon oft nicht eindeutig ist, werden anhand eines **Wortmodells** zunächst *alle* möglichen Zerlegungen ermittelt. Anschließend wird durch **Disambiguierung** der sich ergebenden Varianten anhand eines Rankings die für die Weiterverarbeitung zu verwendende Zerlegung festgelegt. *Beispiel:* Hirnhautentzündung ⇨ Hirnhaut-entzünd-ung (statt Hirn-haut-ent-zünd-ung usw.). Die Festlegung auf eine einzige Variante ist an dieser Stelle sinnvoll; da sowohl die Dokumente des Suchraums als auch die Suchanfragen in gleicher Art zerlegt werden, würde eine Fächerung in Varianten die Distanz zum gesuchten Dokument innerhalb des Suchraums fälschlich erhöhen und andere Dokumente näherliegender erscheinen lassen.

Der Segmentieralgorithmus von MedSearch, der aus Vorverarbeitung und morphologischer Analyse besteht, kommt auch losgelöst vom Gesamtsystem in Anwendungen der linguistischen Analyse und des Retrievals medizinischer Dokumente zum Einsatz ([Schulz 00, Schulz 99] und aktuell im Projekt MORPHOSAURUS, einem mehrsprachigen Wortstammthesaurus zur Unterstützung sprachübergreifenden mehrsprachigen Textretrievals).

Das Wortmodell der morphologischen Analyse

Das Wortmodell der morphologischen Analyse kennt derzeit sieben verschiedene Arten von Morphemen, die bereits oben angesprochen wurden: *Präfixe, Wortstämme* (Grundmorpheme), *Infixe* (Fugenmorpheme), *Suffixe* (aufgeteilt in *Derivationssuffixe* und *Flexionssuffixe*) sowie *Zahlen* und *feste Begriffe*, die nicht mit anderen Morphemen kombinierbar sind. Es entspricht folgendem endlichen Automaten:



Skizze 3.3: Das Wortmodell der morphologischen Analyse

Hinter der Rubrik „feste Begriffe“ verbergen sich eine kleinere Ausnahmeliste von zu unspezifischen Wörtern („Ei“, ...) und Funktionswörtern („und“, „der“ ...), sowie ein langes Verzeichnis an Begriffen der SNOMED und der ICD, deren Zerlegung nicht sinnvoll erschien (Eponyme, Medikamentennamen, sehr seltene Begriffe usw.). Auf diese Weise enthält der Kern des Morphemlexikons einen relativ kleinen Schatz an Morphemen, die großteils häufigen und medizinisch relevanten Begriffseinheiten entsprechen. Diese Kompaktheit gewährleistet eine rasche Zerlegung selbst von längeren Texten und eine gute Auflösbarkeit des größten Teils an medizinischen Fachbegriffen und Wortschöpfungen. – Um eine Vorstellung zu erlangen, in welcher Größenordnung ein Morphemlexikon mit der Menge an analysierten Wörtern anwächst, wurde im Rahmen dieser Arbeit eine kleine Studie durchgeführt, deren Ergebnisse in Kapitel 5.5 sowie in [Schulz 00] näher erläutert sind. Ein Auszug aus dem Morphemlexikon findet sich im Anhang unter Tabelle **(7) Morphemes**.

Um den endlichen Automaten, der das Wortmodell von MedSearch repräsentiert, in der in Skizze 3.3 aufgezeigten einfachen Form zu implementieren, lassen sich die Zustände 1 bis 4 als Funktionen umsetzen, die sich gemäß den durch die Pfeile repräsentierten Konstruktionsregeln rekursiv aufrufen. Jedes Erreichen des Terminalzustandes 4 wird dabei als eine mögliche Zerlegung registriert.

Allerdings sind auch bei einem hypothetischen optimalen Morphemlexikon unzerlegbare Zeichenfolgen schon aufgrund häufiger Rechtschreibfehler und unauflösbarer Abkürzungen sehr frequent. Da Kurzdiagnosen und Arztbriefdiagnosen in aller Regel nur wenig redundante Information enthalten, würde das Verwerfen eines morphologisch nicht komplett zu analysierenden Wortes mit großer Wahrscheinlichkeit zu Fehlindexierungen führen, ein sehr bekann-

tes Problem, das eine generelle Schwachstelle linguistisch motivierter Verfahren gegenüber rein auf Wortähnlichkeiten beruhenden Ansätzen beleuchtet. Deshalb versucht der MedSearch-Algorithmus, in diesen Fällen durch ein Art Backtracking möglichst viel signifikante Information zu „retten“. Dabei wird als Kennzeichen der morphologischen Zerlegung eines Teils des nicht komplett zu analysierenden Wortes gewertet, dass dieser Teil mindestens einen Wortstamm enthält, ggf. inklusive vorangehender und mit folgender Affixe. In diesem Fall wird konsekutiv von links nach rechts nach dem Vorkommen solcher Einheiten gesucht; die jeweils verbleibende Zeichenkette wird iterativ diesem Prozess unterworfen, um weitere Wortstämme oder Komposita mit den zugehörigen Vor- und Nachsilben herauszuarbeiten. Im Gegensatz zu Fällen, in denen eine komplette Analyse möglich ist und in denen der Algorithmus sämtliche Zerlegungen liefert, wird in diesen Ausnahmefällen also eine Forward-Analyse umgesetzt, da die Zahl möglicher inkompletter Zerlegungen bedeutend größer ist. Eine Möglichkeit, bei der vollständigen Berechnung aller Teilzerlegungen deren Anzahl zu reduzieren, wäre die Bedingung der Maximalität, d.h. kein Teil einer Zeichenkette darf nicht unanalysiert bleiben, in dem noch ein Wortstamm steckt. Doch auch bei dieser Bedingung schien der Aufwand zur Umsetzung in keiner Relation zum zu erwartenden sehr geringen Informationsgewinn zu stehen. Die implementierte Forward-Ausnahmebehandlung kann dennoch für die analysierbaren Teile mehrere Zerlegungsmöglichkeiten liefern, unter denen nach den gleichen nachfolgenden Kriterien wie bei der kompletten Zerlegung eine Disambiguierung stattfindet.

Nicht akzeptiert werden vom Algorithmus Teilzerlegungen, die nur Affixe enthalten, da sie alleinstehend zu unspezifisch sind und eher eine Fehlerquelle darstellen. Realistisch für eine Weiterentwicklung wäre hingegen die Möglichkeit, inkomplette Zerlegungen für die folgende Indexierung und das Retrieval zu kennzeichnen, um solche Morpheme als fraglich und daher von geringerem Gewicht zu markieren, ein Ansatz der bisher nicht verwirklicht wurde, da die Komplexität des ohnehin nicht immer leicht durchschaubaren heuristischen Bewertungsmechanismus zu sehr anwachsen würde.

Disambiguierung der möglichen Zerlegungen

Ganz generell folgt auf die morphologische Zerlegung eine Disambiguierung. Sie ist notwendig, weil – vor allem durch eine große Anzahl sehr kurzer Morpheme bedingt – oft mehrere Segmentieralternativen in Frage kommen und in der Regel mehrere, in unserem Fall sogar

sämtliche mit dem Morphemlexikon und dem vorgegebenen Wortmodell dieser möglichen Zerlegungen ermittelt werden. Diese werden durch die Disambiguierung gewichtet und in eine Rangfolge gebracht.

Bei der Disambiguierung gewinnen neben rein formalen Argumenten wie der Anzahl von Morphemen innerhalb einer Zerlegung erstmals im Verlauf des Algorithmus auch semantische Gesichtspunkte eine Rolle, indem jedem Morphem ein Bedeutungsmaß zugeordnet wird. Dabei werden drei verschiedene Gewichte (2, 1 und 0) unterschieden, die nach folgender relativ einfacher Definition den Morphemen des Lexikons attribuiert worden sind:

Gewicht 2: Morpheme, die alleinstehend eine semantische Bedeutung haben (Wortstämme, aber auch Nachsilben wie „-itis“ (Entzündung) usw.);
Gewicht 1: Morpheme, die nur in Verbindung mit anderen Morphemen Bedeutung erlangen (meist, indem sie deren Bedeutung verändern; so die meisten Prä- und Suffixe);
Gewicht 0: Morpheme, die bei der Wortbildung reinen Füllcharakter haben (so alle Infixe und einige Suffixe).

Der in MedSearch umgesetzte Disambiguierungsalgorithmus ist relativ einfach strukturiert. Er vergleicht jeweils zwei Zerlegungen z_1 und z_2 einer Zeichenkette s; dabei werden schrittweise verschiedene Kriterien angewandt, die entweder zum kompletten Ausschluss einer Zerlegungsmöglichkeit führen, die Nachrangigkeit einer der Alternativen feststellen oder, in seltenen Fällen, Gleichwertigkeit bescheinigen. Durch sukzessive Aufnahme der möglichen Zerlegungen in eine gerankte Liste wird eine Reihenfolge ermittelt; der Aufwand für die Vergleiche steigt mit der Anzahl an Zerlegungsmöglichkeiten *n* in der Größenordnung $O(n \log(n))$, doch diese Anzahl ist in aller Regel sehr gering (unter 10, meist 1).

Dabei kommen aktuell 6 Kriterien der Reihe nach zum Einsatz, anhand derer eine Unterscheidung der Zerlegungen angestrebt wird. Sie sind nach ihrer semantischen Relevanz angeordnet, so dass ein untergeordnetes Kriterium nur dann angewandt wird, wenn anhand aller darüberliegenden Gleichrangigkeit der Argumente vorliegt. Im Einzelnen lauten diese 6 Regeln wie folgt:

1) Wenn Zerlegung z_1 eine inkomplette Zerlegung ist (und folglich auch z_2), werden n_1 bzw. n_2 von Zeichen aus s nicht durch Morpheme abgedeckt. Ist n_1 größer als n_2, wird z_1 elimi-

niert, ansonsten z_2. (Mit dieser Regel wird unter inkompletten Zerlegungen die „kompletteste“ ausgewählt.)

2) Wenn an einer Stelle in z_1 (z_2) 4 oder mehr Suffixe aufeinander folgen, in z_2 (z_1) aber an keiner Stelle, wird z_1 (z_2) eliminiert. (Vier Suffixe in Folge sind offensichtlich suspekt.)
3) Wenn an einer Stelle in z_1 3 (z_2) oder mehr Präfixe aufeinander folgen, in z_2 (z_1) aber an keiner Stelle, wird z_1 (z_2) eliminiert. (Drei Präfixe in Folge sind unwahrscheinlich.)
4) Wenn z_1 aus z_2 entsteht, indem ein oder mehrere Morpheme von z_2 in kleinere Morpheme zerlegt werden, wird z_1 eliminiert. Das gleiche gilt umgekehrt, wenn sich z_1 zu z_2 zerlegen lässt. (Dieses Kriterium stellt unter anderem sicher, dass Idiome und feste Begriffe nicht weiter zerlegt werden.)
5) Jedem Morphem des Lexikons ist ein semantisches Gewicht (0, 1 oder 2) zugeordnet. S(z) sei die Summe der Gewichte der Morpheme einer Zerlegung z. Ist $S(z_1)$ größer als $S(z_2)$, so wird z_1 höher gerankt als z_2 und umgekehrt.
6) Morpheme mit semantischem Gewicht 0 werden eliminiert. Dadurch werden nur noch n_1 (n_2) Zeichen aus s durch die entstandenen reduzierten Zerlegungen z_1' (z_2') abgedeckt. Ist n_1 größer als n_2, wird z_1 höher gerankt und umgekehrt.

Beispiel: „Leber transplant -at -ion“ wird „Lebertran -s- plant -at -ion“ vorgezogen, wegen dem Fugen-„s“ von semantischem Gewicht 0 (Kriterium 6).

Zwei morphologische Zerlegungen, die nach jeder der beschriebenen 6 Regeln als gleichwertig einstuft werden, stehen aus Sicht des Algorithmus auf einer Stufe. Eine weitere Differenzierung wäre prinzipiell möglich, die rein formalen Möglichkeiten sind an diesem Punkt allerdings weitgehend erschöpft. Bei einer Weiterentwicklung kämen statistische Komponenten wie eine Berücksichtigung der Häufigkeiten einzelner Morpheme notwendig oder zusätzliche semantische Argumente, etwa die Betrachtung der Morpheme im Zusammenhang. Insgesamt ist es aber sehr selten, dass zwei Zerlegungen gleich bewertet werden müssen und der erforderliche Aufwand für die weitere Abstufung (Untersuchung der Häufigkeiten und Beziehungen von Morphemen in einem repräsentativen Lerndatensatz) wäre vergleichsweise hoch.

Aus bereits erläuterten Gründen wird als Ergebnis der Disambiguierung eine einzelne, die am höchsten gerankte Zerlegung festgesetzt. (Die übrigen Möglichkeiten sowie deren Reihenfolge stehen ggf. aber anderen Applikationen zur Verfügung.) Daher wird bei gleicher Bewertung zweier unterschiedlicher Zerlegungen eine Entscheidung zwischen beiden getroffen, die reproduzierbar sein muss (also etwa nach alphabetischer Reihenfolge). So ist gewährleistet,

dass Dokumente des Suchraumes im Zweifelsfall auf die gleiche Weise wie Suchanfragen zerlegt werden und somit auch subjektiv schlechtere oder falsche Zerlegungen den Zweck der morphologischen Analyse, die innerhalb von MedSearch nur als Teilschritt zur SNOMED-Indexierung und ICD-Klassifikation fungiert, in den meisten Fällen korrekt erfüllen.

Nach Abschluss der morphologischen Analyse arbeitet der weitere Algorithmus nicht mehr mit Morphemen in Form von Zeichenketten, sondern ausschließlich mit den sie repräsentierenden Identifiern.

3.4 SNOMED-Indexierung

Ziel der bisherigen Vorverarbeitung und morphologischen Zerlegung war, die Indexierung von Suchanfragen nun auf einer wesentlich homogeneren terminologischen Basis durchzuführen. Verfahren, die Anfragen direkt auf den Suchraum zu projizieren, haben im Vergleich dazu entweder ein qualitativ schlechteres Ergebnis (mehr Fehlverschlüsselungen, wie etwa bei der Trigramm-Methode), oder sind von der Implementierung her äußerst komplex und imitieren bei näherer Betrachtung häufig diesen Ansatz.

Im nun folgenden Schritt des MedSearch-Algorithmus findet auf Basis der morphologischen Analyse der Suchanfrage eine weitergehende Indexierung mittels des kontrollierten Vokabulars von SNOMED II statt.

Warum ist eine weitere Indexierung überhaupt notwendig? Diese Frage ist zunächst berechtigt. Mit den Identifiern, die als Ergebnis der morphologischen Zerlegung ermittelt wurden, ist der Diagnosentext bereits indexiert; ein Retrieval kann direkt auf Basis dieser Codes erfolgen. Ein solches auf Morphemen, also kleinsten semantischen Einheiten durchgeführtes Retrieval würde aber viele weitergehende Möglichkeiten der Klassifikation anhand von Relationen zwischen medizinischen Begriffen ungenutzt lassen, die SNOMED erschließt. Drei Punkte seien hier explizit genannt, die im MedSearch-Algorithmus Anwendung finden:

- SNOMED enthält eine große Menge an Synonymen. Dadurch kann eine beträchtliche Anzahl von Varianten einer Diagnosenbeschreibung auf der begrifflichen Ebene identifiziert werden. Diese Zwischenebene ist unumgänglich, sofern nicht ein unrealistisch um-

fangreicher Lerndatensatz, der zu jeder Diagnose bzw. Prozedur all ihre denkbaren Umschreibungen enthält, vorausgesetzt werden soll.

- SNOMED ermöglicht die Zuordnung zu semantischen Kategorien (Achsen). Durch die Verschlüsselung ist nicht nur eine Unterscheidung in medizinisch relevante und irrelevante Begriffe möglich, sondern auch in topographische, ätiologische, Krankheitsbezeichnungen usw. entsprechend der SNOMED-Achsen. Dies ist bei der späteren Klassifikation sehr hilfreich (bei einer ICD-Suche wäre ein Krankheitsbegriff wichtiger als die Topographie, bei einer OPS-301-Suche eher eine Prozedur usw.)
- SNOMED enthält Begriffshierarchien. Wenngleich diese in SNOMED II semantisch nicht weiter spezifiziert sind, so ist es doch in vielen Fällen möglich, in der Klassifikation nicht wiedergefundene Begriffe durch einen Oberbegriff zu ersetzen und so die Klasse zu finden, in die der Begriff fällt.

Diese Vorteile, die eine Indexierung medizinischer Dokumente mittels SNOMED für Zwecke der klinischen Klassifikation erbringt, sind bereits durch frühere Arbeiten belegt, so etwa [Ruch 99]. Auch für den Vergleich längerer Texte aus dem medizinischen Bereich mit bereits archivierten Dokumenten (bei uns sind dies lediglich ICD-Diagnosen bzw. OPS-Prozeduren) anhand einer Art Vektorraummodells bietet dieser Indexierungsschritt Vorzüge [Bruijn 97], so zum Beispiel im Bereich der Pathologie, für den die SNOMED aufgrund ihrer historischen Entwicklung gut geeignet ist [Bruijn 96, Bruijn 98]. Daneben sind die bei diesem Schritt anfallenden SNOMED-Indizes an sich als Teilergebnis für andere Applikationen von Interesse und weiter verwendbar.

Ermittlung der SNOMED-Codes

Um den weiteren Weg der Ermittlung der SNOMED-Indizes zu erläutern, folgt eine etwas mathematische Beschreibung des Sachverhalts.

Durch Vorverarbeitung und morphologische Analyse wird eine Funktion p_1 definiert, die eine Projektion aus der Menge aller möglichen Anfragen, d.h. aller endlichen Folgen von Zeichen des endlichen Alphabets, in eine Teilmenge davon, die Menge aller denkbaren Zerlegungen, ermöglicht. Diese Funktion ist surjektiv wie jede Projektion, aber nicht injektiv, da unterschiedliche Dokumente auf identische Zerlegungen projiziert werden können, und daher nicht umkehrbar.

Durch Vorverarbeitung und morphologische Analyse wird des Weiteren eine Funktion p_2 definiert, die die Menge aller einem SNOMED-Code zugeordneten Mengen von Dokumenten (oft entspricht ein SNOMED-Code ja einer ganzen Reihe von synonymen Begriffen) auf die Menge ihrer Zerlegungen projiziert (genauer, eine Menge von Mengen von Zerlegungen). Diese Funktion ist wiederum per Definition surjektiv. Da im Folgenden über eine noch zu definierende *best fit*-Abbildung die Verschlüsselung des Suchbegriffs auf der Ebene der Zerlegungen stattfinden soll, ist für die Funktion p_2 aber eine Umkehrbarkeit, d.h. Injektivität und damit Bijektivität, erforderlich. Nur so kann über p_2^{-1} dem zerlegten Suchbegriff schließlich eine Menge von SNOMED-Codes zugeordnet werden, da nur dann das Diagramm 3.6, das diesen Sachverhalt illustrieren soll, kommutiert.

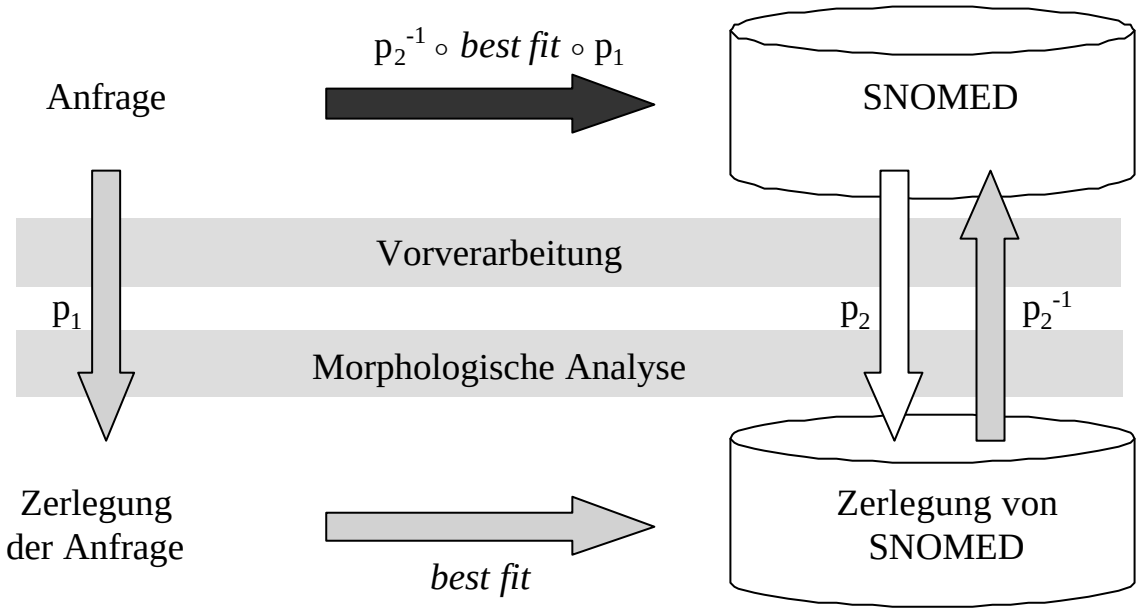


Diagramm 3.4: Prinzip der Ermittlung der SNOMED-Codes

Wie Diagramm 3.4 aufzeigt, handelt es sich bei der SNOMED-Indexierung mittels einer *best fit*-Funktion im Prinzip bereits um ein Retrieval, bei dem die durch die Identifier der morphologischen Analyse indexierte Suchanfrage in einem auf die gleiche Art indexierten Suchraum, den Folgen zerlegter SNOMED-Terme, wiederzufinden ist. Aus diesem Grund gleicht Diagramm 3.4 dem in der Folge das Prinzip des MedSearch-Retrievals illustrierenden Diagramm 4.1, wobei dort die Definition der eigentlichen *best fit*-Retrievalfunktion natürlich eine völlig andere ist.

Da für die Abbildung der zerlegten Suchanfrage in den Suchraum eine morphologische Analyse der gesamten SNOMED II Voraussetzung ist, die natürlich vorab stattfinden muss, ist vor der Umsetzung des Algorithmus ein beträchtlicher Vorbereitungsaufwand notwendig. Schwerwiegender noch ist die Tatsache, dass bei späteren Verbesserungen des Morphemlexikons (Hinzunahme und Löschen von Morphemen sind sehr häufig erforderlich) oder auch des Abkürzungsverzeichnisses kein einfacher Backtracking-Mechanismus o.ä. existiert, mit dem von der Änderung betroffene SNOMED-Termini herausgefiltert werden könnten und jeweils die Korrektheit der Zerlegungsliste dadurch gewährleistet werden muss, dass eine komplette Neuzerlegung von SNOMED stattfindet. Dies wiederum impliziert in der Folge eine Neusegmentierung und -indexierung des Suchraums für das MedSearch-Retrieval (ICD- sowie OPS-301-Lerndatensatz). Daher ist anzuraten, erforderliche Korrekturen am Morphemlexikon anzusammeln und en bloc durchzuführen, doch auch dann bleibt der Rechen- und Zeitaufwand jeweils beträchtlich. – Wie die morphologische Vorabanalyse einzelner SNOMED-Termini aussieht kann der Tabelle **(8) SNOMED** entnommen werden.

Ein weiteres Problem bei der Umsetzung des Algorithmus bildet die Injektivität der Funktion p_2, die durchaus nicht selbstverständlich ist. Daher ist bereits bei Erstellung der Regeln für die morphologische Analyse und insbesondere beim Sammeln der Datensätze, vor allem des Morphemlexikons, darauf zu achten, dass die Möglichkeit weitgehendst ausgeschlossen wird, zwei zu unterschiedlichen Codes gehörende SNOMED-Einträge auf eine identische Zerlegung zu projizieren. In der Praxis muss die Eindeutigkeit der Zerlegungen dadurch gewährleistet werden, dass die Zerlegung der SNOMED bei jeder Neuberechnung nach identischen Elementen durchsucht und ggf. die Ursache behoben wird. Da in der Regel ein bestimmtes Morphem des Morphemlexikons diese Ursache darstellt, impliziert das eine erneute Änderung dieses Datensatzes mit der Folge einer abermals notwendigen Zerlegung des SNOMED.

Ein eindrückliches Beispiel, welche Fehler schon durch nur ein falsch eingeschätztes Morphem auftreten können, ist die Gewichtung des Funktionsworts „weil“ mit dem Wert 0 (medizinisch nicht relevant). Dadurch wird der Begriff „weil“ auf die leere Zerlegung projiziert, und als Folge davon die Diagnose „Morbus Weil“ auf die Zerlegung „morbus“, was der Diagnose „Krankheit“ entspräche. Damit ist aber nicht mehr definiert, ob durch die Umkehrung der Funktion p_2 nun „morbus“ auf „**F00102** Morbus“ oder „**D01970** Morbus Weil“ abgebildet werden soll, so dass bei einem Dokument, das den (häufigen!) Begriff „Krankheit“ enthält, nun bei der Indexierung regelmäßig der Schlüssel „**D01970** Leptospirosis icterohaemorrhagi-

cae“ codiert wird. – Um Fehler dieser Art zu vermeiden, ist daher stets darauf zu achten, dass Projektion und Zerlegung, eingeschränkt auf den Suchraum (nicht auf die gesamte Grundmenge!) injektiv sind.

Abbildung des zerlegten Dokuments auf den nächstgelegenen Eintrag in der Zerlegung von SNOMED (*best fit*-Funktion)

Bislang wurde das Problem der Indexierung eines Dokuments auf eine einfachere, standardisierte Ebene abgebildet und der Rückweg beschrieben. Nun bleibt die eigentliche Aufgabe, die Definition der *best fit*-Funktion, mit der ein zerlegtes Dokument auf den nächstgelegenen Eintrag in der Menge aller möglichen Folgen von Zerlegungen von SNOMED-Texten abgebildet werden soll.

Ein möglicher Ansatz wäre, für jedes zerlegte Dokument sukzessive die Liste an bereits im Vorfeld segmentierten SNOMED-Einträgen zu überprüfen, dann den längsten in diesem Dokument enthaltenen Eintrag zu wählen, die korrespondierenden Morpheme aus dem Dokument zu streichen und den Rest wiederum iterativ mit der Liste an zerlegten SNOMED-Einträgen durchzugehen, bis schließlich die Menge der restlichen aus der Zerlegung des Dokuments stammenden Morpheme leer ist oder nur noch Elemente enthält, die nicht mehr durch SNOMED-Einträge abgedeckt werden. (Alternativ könnte nicht nur für den längsten, sondern für jeden enthaltenen Eintrag so vorgegangen werden, wobei am Ende wieder eine Disambiguierung stattfinden müsste.) Eine solche Vorgehensweise ist aber aus Gründen der Rechenzeit nicht praktikabel.

Dieser hypothetische Ansatz lässt sich jedoch verfeinern, indem man für jedes Morphem des Morphemlexikons eine Liste aller SNOMED-Einträge erstellt, in deren Zerlegung es enthalten ist. So kann ohne großen Aufwand der weiteste Teil des Suchraums von vornherein ausgeschlossen werden, da er zu dem Anfragedokument disjunkt ist.

Auch bei diesem Ansatz ist jedoch für die gängigsten Morpheme der Teilraum von SNOMED, in dem sie enthalten sind, noch beträchtlich – allein 4936 Eintrage der SNOMED II enthalten beispielsweise das Morphem „syndrom“. Darum ist die Idee des jetzigen Ansatzes weitergehend. Dabei wird zur Voraussetzung erklärt, dass nur dann ein SNOMED-Eintrag als in einem Dokument enthalten registriert werden soll, wenn mindestens alle eine eigene Be-

deutung tragende Morpheme (d.h. alle Morpheme von Gewicht 2, im Folgenden auch als *Lexeme* bezeichnet) des zerlegten Eintrags in der Zerlegung des Dokuments enthalten sind.

Damit genügt es nun, aus jeder Zerlegung eines SNOMED-Eintrags **ein** Lexem herauszuschreiben, das somit als Zeiger fungiert. Dabei werden diese Zeiger möglichst gleichmäßig verteilt, so dass nun auch ein Lexem wie „krank“ (Krankheit) oder „-itis“ (Entzündung) auf maximal etwa 40 Einträge verweist, in denen es vorkommt. (Die übrigen sind dann nur über andere Schlüssellexeme wiederzufinden.) Das ergibt bei der Eingabe einer Kurzdiagnose anstatt bis zu mehreren Tausend in der Regel nur noch 50 bis 150 SNOMED-Einträge, die überhaupt noch betrachtet werden müssen.

Nachteil bei dieser Methode ist, dass eine „unscharfe“ Suche nach Einträgen, von denen nur „fast“ alle Lexeme im Suchausdruck vorkommen, nicht möglich ist. Diese ist aber in Fällen, bei denen die normale Indexierung nichts liefert, ohnehin kaum angebracht, da sie zu fehlerbehaftet ist.

Neben den Bedeutungstragenden Morphemen werden in Zweifelsfällen auch die bedeutungsmodifizierenden (Gewicht 1) überprüft. So kann nun unterschieden werden zwischen prim-är, Prim-at, Prim-id-on usw. Ähnlich wie in der morphologischen Zerlegung ist dazu ein endlicher Automat implementiert worden, der nicht nur die *Longest Match*-Variante berechnet, sondern sämtliche Indexiermöglichkeiten betrachtet und im Einzelnen gewichtet.

Zu berücksichtigen ist bei dieser Gewichtung, dass die vorige Bearbeitung keineswegs eine Menge von Wortstämmen, Affixen usw. liefert, sondern dass Wortgrenzen (durch Klammerung) und die Reihenfolge der Wörter erhalten bleiben. Diese Information wird hier ausgewertet. Negativ ins Gewicht fallen dabei:

1) Jedes Herauslösen eines Lexems aus einem Wort.
2) Jedes Zusammenfassen von Lexemen aus verschiedenen Wörtern.
3) Das Streichen bedeutungsmodifizierender Morpheme.
4) Das Hinzufügen bedeutungsmodifizierender Morpheme.
5) Nicht berücksichtigt werden bisher die Abstände, die beim Zusammenfassen bzw. Herauslösen überwunden werden müssen.

6) Jedes Wort muss Lexeme beinhalten. Überall, wo alle Lexeme „wegverschlüsselt“ wurden, werden die verbleibenden Morpheme des Wortes umgehend gestrichen, was wie beschrieben negativ ins Gewicht fällt.

Auf diese Art ergibt sich in den meisten Fällen die subjektiv beste Verschlüsselungsmöglichkeit als erste Wahl. Um auch hier Aufwand und Komplexität in Grenzen zu halten sowie eine identische Zerlegung von Dokumentenkollektion und Suchanfrage zu gewährleisten, wird wie bei der Disambiguierung morphologischer Zerlegungen nur die im Ranking an erster Stelle stehende Möglichkeit weitergegeben.

Bis hier wurde im Wesentlichen der Ansatz, der [Brigl 94, Brigl 95] zu Grunde liegt, verfolgt und verfeinert. Nicht mit in die Indexsuche gehen bisher die in der Vorverarbeitung herausgefilterten Informationsqualifikatoren ein. Ansonsten kann die SNOMED-Indexierung hauptsächlich durch Optimieren des zu Grunde liegenden Morphemlexikons qualitativ verbessert werden. Wie für die Vorverarbeitung und die morphologische Analyse gilt aber auch hier, dass falsche Resultate oft in der nächsten Stufe wieder aufgefangen werden, da letztendlich anhand des ICD- bzw. OPS-301-Lerndatensatzes klassifiziert wird, dessen Einträge mit dem gleichen Algorithmus in der identischen Weise richtig oder falsch zerlegt werden.

Die zusätzliche L-Achse für nicht nach SNOMED indexierbare Lexeme

In einer großen Studie wurden 1995 4 typische Fehlerquellen bei der automatischen SNOMED II-Indexierung unterschieden [Carter 96]. 65 % der ermittelten Fehler fielen dabei in eine einzige Kategorie, bei der in SNOMED enthaltene Termini durch Varianten in der individuellen Terminologie des Benutzers nicht erkannt wurden. (Zudem betrachtet die Studie ausschließlich pathologische Berichte, also Texte, bei denen der Wortschatz im Verhältnis zum gesamten Fachgebiet Medizin zusätzlich eingeschränkt ist.) Nachrangige Fehlerquellen waren: Nichtcodierung wegen trotz medizinischer Relevanz nicht vorhandenen SNOMED-Eintrags; inkorrekt zugeordnete SNOMED-Indizes; überflüssige oder bedeutungslose Codes (vor allem dort, wo SNOMED nicht wirklich disjunkt ist, also speziell bei der D-Achse, die Krankheitsbezeichnungen enthält, welche oft durch Begriffe aus anderen Achsen zusammengesetzt sind).

In einer weiteren Studie, bei der ein unserem Ansatz vergleichbarer Algorithmus (MedLEE) die automatische Indexierung von Krankengeschichten mit SNOMED RT umsetzte, wurde festgestellt, dass 15 % der klinisch als signifikant erachteten Information eine den Wortschatz SNOMEDs überschreitende Terminologie benötigte [Lussier 01]. Aus all diesen Gründen heraus erscheint es dringend geboten, sich bei der Indexierung von Dokumenten trotz aller Vorteile der SNOMED nicht auf sie zu beschränken. Ziel ist nicht, dem Benutzer ein kontrolliertes Vokabular vorzuschreiben, sondern vielmehr anhand der auftretenden Begriffe eine Kontrolle der Nomenklatur.

Eine solche Erweiterung der Nomenklatur, die im Grunde dynamisch sein sollte (indem sie neu auftretende Begriffe registriert und erlernt), wird vom Algorithmus durch die Hinzunahme einer Tabelle mit in SNOMED nicht vorhandenen Einträgen umgesetzt. In diese Tabelle, die als *L-Achse* bezeichnet werden soll (in Analogie zu den 7 Achsen von SNOMED II), werden alle Lexeme aufgenommen, die anhand der SNOMED bei der Indexierung des Suchraums (ICD- sowie OPS-301-Lerndatensatz) nicht zu verschlüsseln sind. Fast alle dieser Lexeme sind medizinisch relevant. Sie werden in der L-Achse, die im Gegensatz zu den Achsen SNOMEDs nicht hierarchisch angeordnet ist und keine semantische Einheit bildet, durch ein „L“, gefolgt von ihrem fünfstelligen Morphemidentifier codiert. Morpheme von Gewicht 0 oder 1, die bei der Indexierung des Suchraums unverwendet bleiben, werden hingegen verworfen.

Diese zusätzliche Achse stellt gewissermaßen eine Umgehung der SNOMED dar für Krankheits- oder Prozedurbezeichnungen, die mit SNOMED nicht komplett indexierbar sind, aber im ICD- bzw. OPS-301-Lerndatensatz enthalten. Sie können so trotz der Probleme bei ihrer Indexierung korrekt im Suchraum wiedergefunden werden. Allerdings bedingt dieses Konzept in der Folge wiederum einige Ausnahmeregelungen für den Retrievalalgorithmus.

3.5 Weitere Synonymverarbeitung und verwandte Begriffe

Ein zentraler Punkt jeder automatischen Indexierung von Texten ist das Identifizieren synonymer Begriffe. Dies wird innerhalb des MedSearch-Algorithmus weitestgehend durch die große Zahl synonymer Einträge, die in der SNOMED verzeichnet sind, gewährleistet. Dennoch ist eine zusätzliche Synonymbehandlung, wie sich auch experimentell ergab, sehr wünschenswert.

Einer zu weitgehenden Synonymgleichsetzung stellt sich indes das Problem der sich stetig vergröbernden Granularität der Begriffseinheiten entgegen. In vielen Fällen von Sinnverwandtheit sind Begriff und Synonym nur bedingt deckungsgleich, so dass ihre Identifikation eine Klassenbildung im Kleinen vorwegnimmt.

Eine begrenzte zusätzliche Liste an Synonymen ist aber durchaus sinnvoll, da viele Zuordnungen nur so zu lösen sind. So kommt es vor, dass zwei Lexeme für die gleiche morphologische Einheit stehen, wie zum Beispiel in „infiz-iert“ und „Infekt-ion“. In diesem Fall wird der zweite Begriff in SNOMED wiedergefunden, der erste aber nicht (auch nicht in der dortigen Synonymliste) und daher auf die L-Achse abgebildet. Durch den Algorithmus werden derzeit die nach SNOMED indexierten Elemente der Tabelle **(9) Synonymes** (soweit deren Indexierung nicht ohnehin übereinstimmend ist) identifiziert.

Ein Problem hierbei stellt die Frage dar, welches Maß an Übereinstimmung vorliegen muss, um ein Wort in der Synonymliste wiederzufinden. Können in obigem Beispiel nun automatisch auch „infiz-ieren“ und „Infekt-ion“ gleichgesetzt werden? Aktuell fordert der Algorithmus die Übereinstimmung sämtlicher Lexeme *und* der bedeutungsmodifizierenden Morpheme. Andere Möglichkeiten wären, die weitere Verarbeitung in mehrere Varianten aufzuteilen und zu gewichten, was aber Zeitaufwand und Komplexität des Systems erhöht, oder im Sinne eines dynamischen Modells bei problematischem Retrieval im nächsten Abschnitt auf diese Stufe zurückzukehren, um die Bedingungen etwas zu „lockern“.

Eine mögliche Variante wäre auch die Synonymerfassung bereits auf Ebene der Lexeme [Schulz 99]. Allerdings würde hier eine Abbildung auf Lexeme gleicher Bedeutung in vielen Fällen ein risikoreiches Unterfangen darstellen; da einzelne Lexeme oft nur im Zusammenhang mit bedeutungsmodifizierenden Morphemen oder im Verbund mit weiteren Lexemen ihre volle Bedeutung entfalten, kann es zu groben oder fälschlichen Verallgemeinerungen kommen. Des weiteren würde in diesem Fall die Streichung oder Weiterverwendung verbleibender bedeutungsmodifizierender Morpheme, die bei der SNOMED-Indexierung eine Rolle spielen, zusätzliche Ambiguitäten produzieren.

Für die Behandlung von Begriffen, die nicht identisch sind, aber in einer semantischen Relation stehen, wäre generell ein komplexerer Ansatz und die Verwendung eines Thesaurus notwendig. Was derzeit indes bereits umgesetzt wird, ist das in SNOMED II implementierte

Konzept der verwandten Begriffe (*related terms*), indem bestimmte Begriffe wie „Hepatitis“ auf die durch SNOMED angegebene Menge von Termini (in diesem Fall „Leber“ und „Entzündung“) abgebildet werden. Dadurch wird eine zusätzliche Identifikation mit alternativen Diagnosenbezeichnern (hier „Entzündung der Leber“) möglich, sofern dies nicht schon auf Ebene der Morpheme geschieht.

Generell sollte eine Gleichsetzung von Synonymen aber zurückhaltend gehandhabt werden. Auch auf Basis einer nur teilweise korrekten SNOMED-Indexierung kann eine korrekte Klassifikation noch erfolgen, da die Diagnosen bzw. Prozeduren des ICD- und OPS-301-Lerndatensatzes auf die gleiche Weise indexiert sind. Eine zu großzügige Synonymgleichsetzung hingegen kann durch das Retrieval, das im nun folgenden Kapitel beschrieben wird, nicht mehr aufgelöst werden.

4 Methodik des Retrievals der SNOMED-indexierten Suchanfrage

Ziel des sogenannten MedSearch-Retrievals ist die vollautomatische ICD-Klassifikation von Diagnosen und OPS-301-Klassifikation von Prozeduren. Es operiert auf Basis der vorangegangenen Indexierung von Suchanfrage und Dokumentenkollektion (dem ICD- bzw. OPS-301-Lerndatensatz, vgl. Abschnitt 5.3) mittels SNOMED II.

Systeme zur automatischen Klassifikation finden seit Jahren verbreitet Anwendung. Bereits früh existierten Algorithmen zur teilautomatischen ICD-Klassifikation wie in [Michel 95], bei denen der Benutzer zwischen den Endergebnissen und z.T. auf Zwischenstufen über eine Art Browser manuell eine Auswahl unter verschiedenen Alternativen treffen muss. Inzwischen finden sich bereits Ansätze, die ICD-Verschlüsselung direkt aus der elektronischen Krankenakte heraus zu realisieren [Blanquet 99], die in zunehmendem Maße verbreitet ist. Eine korrekte Klassifikation ohne jede ärztliche Überprüfung bleibt allerdings bis auf weiteres ein unrealistisches Ziel, wiewohl solche Verfahren für statistische Anwendungen ihren Zweck erfüllen.

Wie bereits auf Ebene der Morpheme möglich, kann ein einfacher mathematischer Ansatz wie das Vektorraumretrieval nun auch direkt auf Basis der SNOMED-indexierten Suchanfrage erfolgen. Diese Möglichkeit wurde zum Zweck des Vergleichs des MedSearch-Algorithmus mit einem alternativen Ansatz herangezogen (Kapitel 6 und [Franz 00]). Das vergleichsweise komplexere MedSearch-Retrieval hingegen macht in mehrfacher Hinsicht von semantischen Aspekten der SNOMED-Indexierung Gebrauch. So kann anhand der vorangegangenen Indexierung jeder aufgefundene Begriff einer der SNOMED-Achsen zugeordnet werden (von Elementen der L-Achse abgesehen), der damit eindeutig einer semantischen Kategorie (Topographie / Krankheiten / Prozeduren usw.) zugeordnet werden kann. Dieses Wissen erlaubt es, gezielt nach Begriffen zu suchen, etwa nach Krankheiten bei der ICD-Klassifikation oder nach Prozeduren bei der Suche im OPS 301. Beispielsweise wird bei der Klassifikation der Suchanfrage „Strumektomie“ mittels OPS-301 vorrangig nach dem Begriff „Ektomie“ (Prozedur-Achse) und nachrangig nach „Struma“ (Topographie) zu suchen sein. – Ein weiterer Vorteil der SNOMED-Indexierung ist die Möglichkeit, sich die implizite Hierarchie der SNOMED II zu Nutze zu machen, durch die SNOMED-Termini in einer Art Klassifikationsschritt zu Oberbegriffen zusammengefasst werden können.

4.1 Prinzip der Ermittlung der ICD-Codes

Die Situation vor dem letzten noch fehlenden Schritt zur ICD-9, ICD-10 bzw. OPS 301-Klassifikation der Suchanfrage, dem MedSearch-Retrieval, ist in Diagramm 4.1 wiedergegeben. Sie ähnelt im Wesentlichen den Verhältnissen vor der letzten Stufe der SNOMED-Indexierung. – Soweit nicht anders erwähnt, soll im Folgenden „ICD“ stellvertretend für ICD-9, ICD-10 oder auch den Prozedurkatalog OPS 301 stehen.

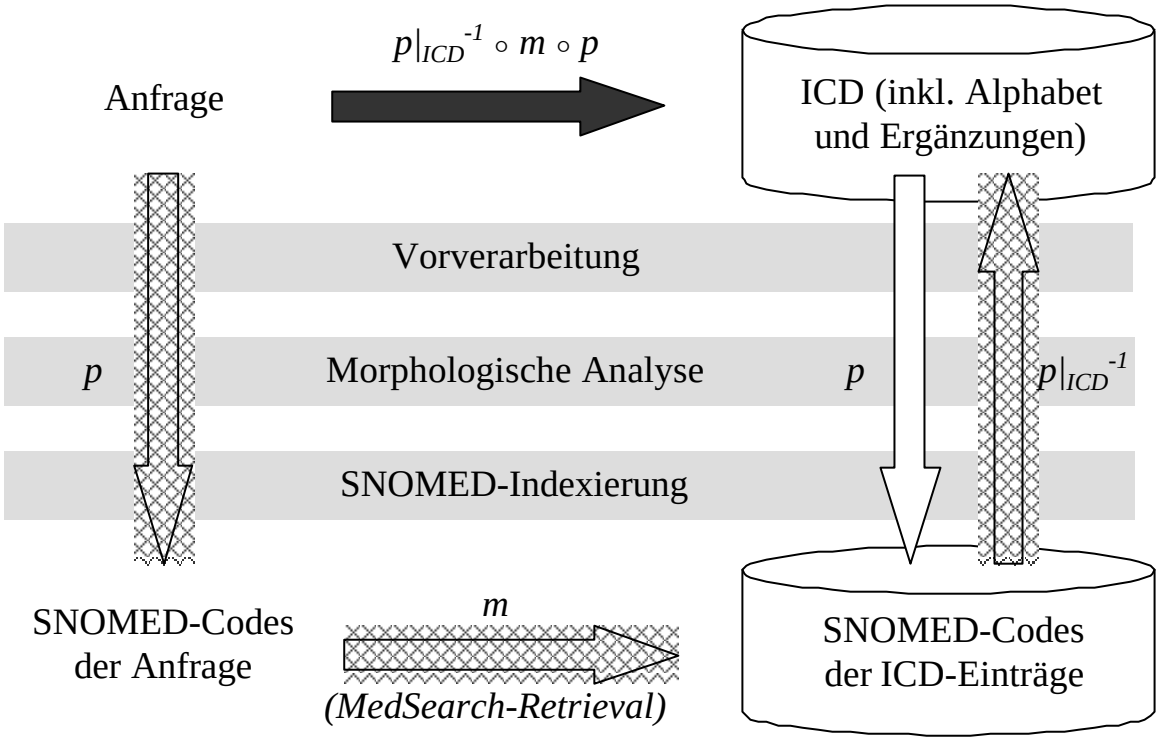


Diagramm 4.1: Prinzip der Ermittlung der ICD-Codes

Wiederum wird durch die bisherigen Schritte des Algorithmus eine Projektion p von der Menge aller möglichen Ausgangsdokumente auf die Menge aller endlichen Folgen von SNOMED-Codes definiert. Da SNOMED-Codes innerhalb einer Indexierung mehrfach vorkommen können, kann der Begriff „Folgen“ nicht durch „Mengen“ ersetzt werden, obwohl es in diesem Fall auf die Reihenfolge innerhalb der Folgen nicht ankommt. Sie ist zum einen oft nicht eindeutig festzustellen, da die Morpheme, denen ein SNOMED-Code zugeordnet wird, nicht unbedingt zusammenhängend in der Zerlegung des Dokuments vorliegen. Zum anderen ist die semantische Bedeutung dieser Reihenfolge beim Betrachten relativ kurzer Texte nach-

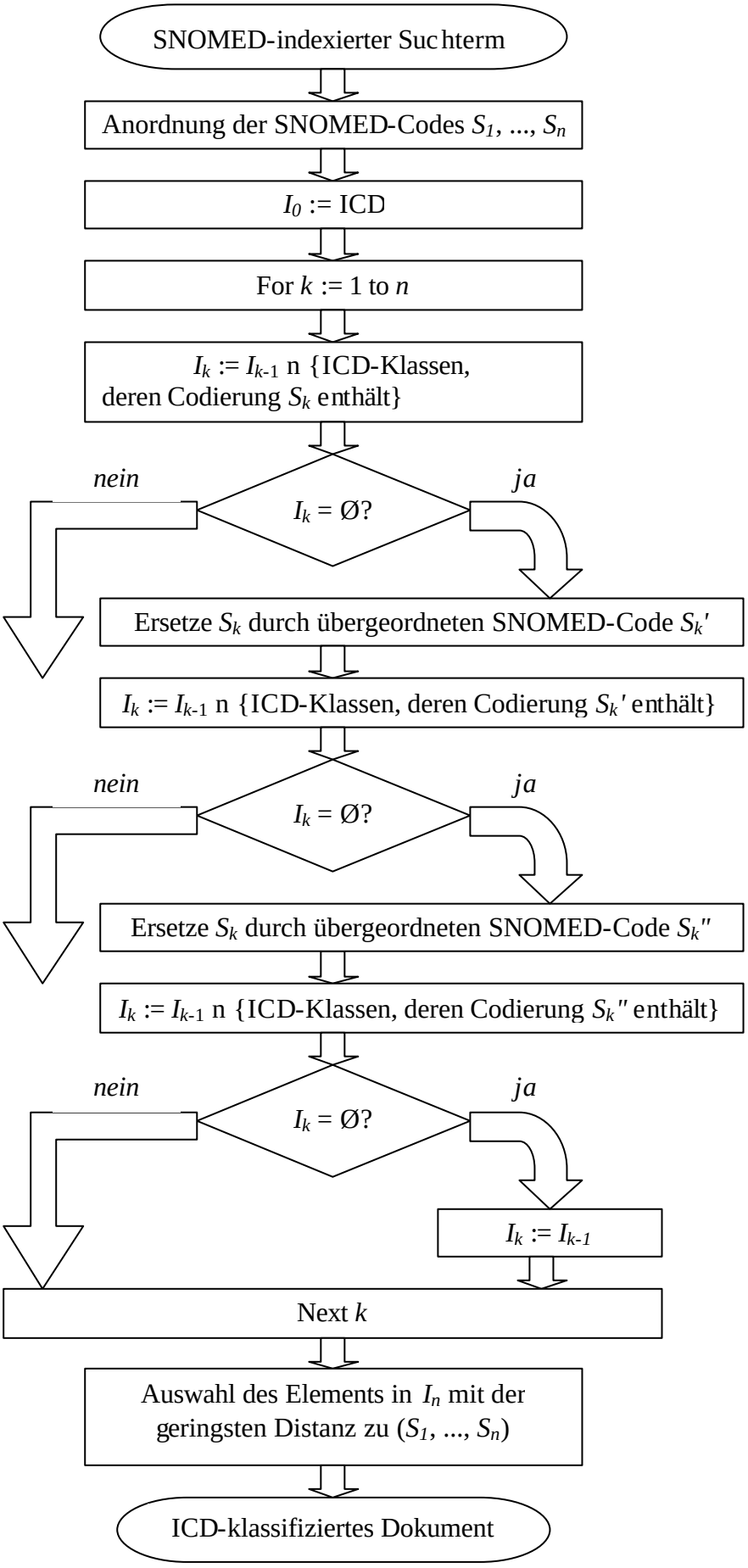
rangig, so dass im MedSearch-Retrieval die Reihenfolge der SNOMED-Codes innerhalb der Folge anhand anderer Kriterien definiert wird (siehe Abschnitt 4.3).

Wie in 3.4 ist auch hier die Projektion p nicht injektiv, und daher nicht umkehrbar. Die Umkehrbarkeit der auf die Grundmenge der ICD-Klassen (immer einschließlich der alphabetischen Einträge und der Ergänzungsliste) beschränkten Projektion $p|_{ICD}$ jedoch muss wiederum gewährleistet sein, d.h. wie in 3.4 ist darauf zu achten, dass die Projektionen den ICD-Klassen umkehrbar eindeutig zugeordnet sind, um die Klassen später korrekt attribuieren zu können, die in den Tabellen **(10) ICD-9**, **ICD-10** und **OPS-301** verzeichnet sind. Dies muss erneut dadurch sichergestellt werden, dass nach der SNOMED-Indexierung des ICD-Lerndatensatzes **(11) ICD-9_ABC, ICD-10_ABC** bzw. **OPS-301_ABC**, die abermals bereits im Vorfeld erfolgen muss, um den Suchraum für das MedSearch-Retrieval bereitzustellen, die Injektivität bzw. Bijektivität (auf der Zielmenge) der Funktion $p|_{ICD}$ überprüft wird, d.h. es dürfen nur dann identische Folgen (unabhängig von der Reihenfolge) von SNOMED-Codes zwei Diagnosen repräsentieren, wenn diese derselben ICD-Klasse zugeordnet sind. – Diese Einschränkung spielt hier allerdings eine geringere Rolle als in 3.4, da eine qualitativ hochwertige SNOMED-Indexierung und eine eindeutige Klassenzuordnung des Lerndatensatzes aus ICD-Systematik, Ergänzungsliste und Alphabet die Eindeutigkeit und damit Reversibilität der ICD-Verschlüsselung weitestgehend garantieren.

Was der letzten Stufe des Algorithmus, dem eigentlichen MedSearch-Retrieval, übergeben wird, sind nun lediglich SNOMED-Identifier (inklusive L-Achse). Alle ICD-klassifizierten Diagnosen, mit denen verglichen wird, liegen im Suchraum in gleicher Form mittels SNOMED II präcodiert vor. Zu berücksichtigen ist wiederum, dass bereits eine unbedeutende Änderung der Lexika oder des Algorithmus die Neuzerlegung des gesamten Lerndatensatzes, der dem Suchraum zu Grunde liegt (ICD-Systematik + alphabetisches Verzeichnis + Ergänzungen) notwendig macht, um ein korrektes Retrieval zu gewähren.

4.2 Überblick über das MedSearch-Retrieval

Das eigentliche MedSearch-Retrieval ist in mehrere Einzelschritte aufgeteilt, über deren Zusammenhang das folgende Flussdiagramm 4.2 einen Überblick vermitteln soll und das im Anschluss kurz erläutert wird.



Flussdiagramm 4.2: MedSearch-Retrieval

Zunächst werden im Verlauf des in Diagramm 4.2 skizzierten MedSearch-Retrievals die SNOMED-Codes (immer inklusive der L-Achse) innerhalb der dem Dokument entsprechenden Schlüsselfolge in bestimmter Weise gewichtet bzw. angeordnet (4.3). Im Folgenden wird versucht, eine Menge von Elementen im Suchraum (dem auf gleiche Weise codierten ICD-Lerndatensatz) zu finden, die dieser Schlüsselfolge möglichst nah liegen, indem die Menge an ICD-Klassen, in deren Verschlüsselung der in der Reihenfolge erste SNOMED-Code enthalten ist, geschnitten wird mit der dem zweiten Code über $p|_{ICD}^{-1}$ zuzuordnenden ICD-Klassen, diese Schnittmenge dann wiederum mit der des dritten usw. (4.4). Sollte die Schnittmenge bei einem bestimmten Code leer sein, wird versucht, für den diesem Code zugehörigen SNOMED-Eintrag einen Oberbegriff zu finden (4.5). Die Elemente der verbleibenden ICD-Teilmenge werden schließlich einzeln auf ihre Distanz zum indexierten Suchdokument überprüft und das nach einer zu definierenden Metrik nächste Ergebnis ermittelt (4.6).

4.3 Anordnung der SNOMED-Codes

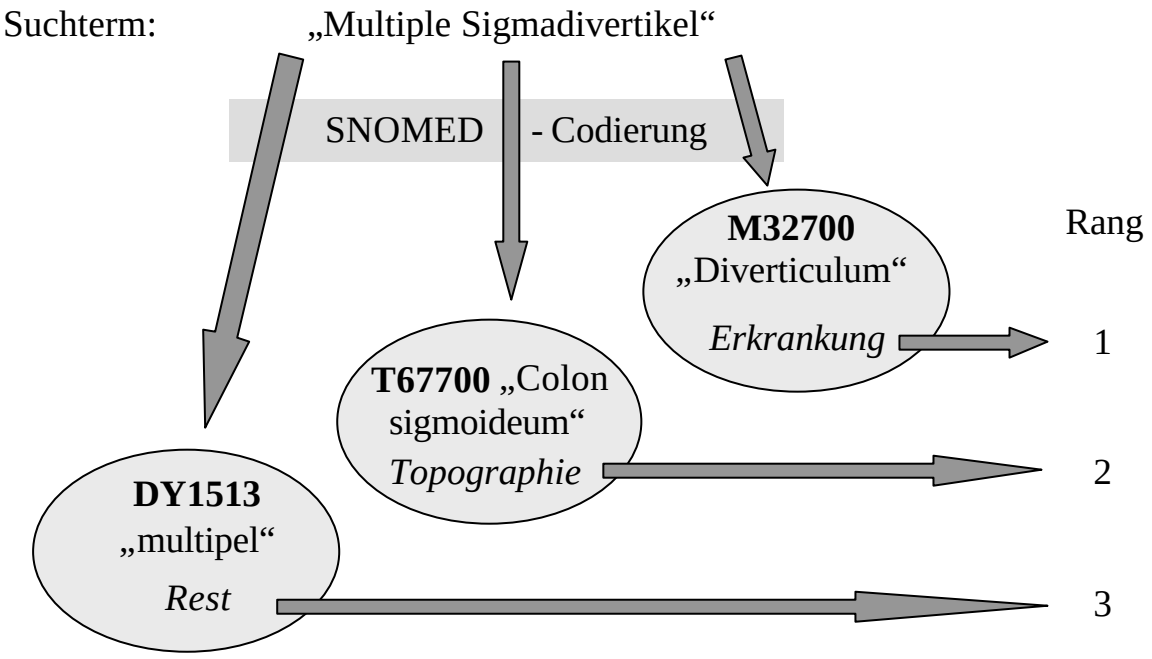
Durch die bisherigen Schritte bis zur SNOMED-Indexierung wird dem MedSearch-Retrieval eine Reihe von SNOMED-Codes übergeben, der nun eine ICD-Klasse zugeordnet werden soll. Dabei enthält das Suchdokument oft eine Fülle unterschiedlicher Informationen, die für die Klassifikation zum Teil relevant, zum Teil aber nur bedingt wichtig bis unbedeutend sind. Entsprechend werden die SNOMED-Codes, die diesen Informationen entsprechen, zunächst gewichtet, bzw. nach ihrer Relevanz angeordnet (gerankt).

Das Sortieren der SNOMED-Codes einschließlich der L-Achse findet nach folgenden 4 Regeln statt:

1) **Krankheitsbezeichner** stehen an erster Stelle. Unter „Krankheitsbezeichner“ im weiteren Sinne fallen die SNOMED II-Einträge D00000-DY0YYY, DY1400-DY10YY, DY2000-DYYYYY, F00000-FXYYYY, FYY000-FYYYYY, M00000-M492YY, M49400-M9YYYY und MX0300-MYYYYY.
   Im Falle der Klassifikation in den OPS 301-Katalog stehen stattdessen **Prozeduren** an der ersten Stelle. Dazu zählen die Einträge P00000 - PYYYYY
2) **Topographische Begriffe** stehen an zweiter Stelle. Dazu zählen die Termini unter T00400-T005YY und T01000-TY9YYY.

3) **Alle übrigen Begriffe,** also ätiologischer Art, Berufe usw. stehen an letzter Stelle, mit Ausnahme von Elementen der L-Achse.
4) **Lexeme** (Elemente der L-Achse) werden der höchsten Klasse zugeteilt, die leer bleibt. Das heißt: Sind keine Codes vorhanden, die auf einen Krankheitsbezeichner hinweisen, aber ein (unbekanntes) Lexem, so wird dieses als potenzieller Krankheitsbegriff gewertet und an die erste Stelle gesetzt. Sind aber alle Arten von Begriffen vorhanden, stehen restliche Lexeme in der Suchreihenfolge zuletzt.

Das Beispiel 4.3 illustriert diese Regeln anhand des Suchbegriffs „Multiple Sigmadivertikel".



Beispiel 4.3: Ranking von SNOMED-Codes nach ihrer semantischen Bedeutung

Die Gewichtung der in der Verschlüsselung des Dokuments vorhandenen SNOMED-Codes hat essenzielle **Vorteile.** Durch die Typisierung der Codes anhand der Einteilung im Wesentlichen entlang der SNOMED-Achsen in die Kategorien „Erkrankung", „Topographie" und „Rest" ist eine grobe Semantik verfügbar. Mit dieser Semantik können Begriffe, die für ICD-Klassen Schlüsselfunktion besitzen (also im Wesentlichen solche, die Krankheiten bezeichnen) vorrangig gesucht werden; Zusätze werden zweitrangig.

Auf diese Art wird der Suchraum, beginnend mit dem ganzen ICD, möglichst schonend eingeschränkt, d.h. die Suche kann auf das Umfeld der wesentlichen Begriffe – im Beispiel 4.3 etwa des Begriffs „Divertikulum“ – fokussiert werden. Eine Einschränkung des Suchraums aber ist aus praktischen Gründen sinnvoll, da ein komplettes Durchsuchen aller Klassen, deren SNOMED-Indexierung mit der des Dokuments nicht disjunkt ist, aufgrund eines umfangreichen Beispiellexikons und der zum Teil sehr langen Diagnosen, die oft auch für die Klassifikation Unwesentliches enthalten, selbst mit schnellen Rechnern in Echtzeit kaum zu bewältigen ist.

Des Weiteren wird durch diese Gewichtung der SNOMED-Codes auch in Fällen, in denen die korrekte Klasse nicht aufgefunden wird, doch oft zumindest eine Klasse mit dem korrekten Krankheitsbegriff ermittelt. Die Fehlklassifikation „ICD-9 **412** alter Myokardinfarkt“ an Stelle von „ICD-9 **410** akuter Myokardinfarkt“, die sich an dem Krankheitsbegriff „**D72560** Myokardinfarkt“ orientiert, läge der korrekten Klasse sicher semantisch näher als etwa „ICD-9 **204.0** Akute lymphatische Leukämie“ bei vorrangiger Suche nach dem SNOMED-Eintrag „**MX0104** akut“.

Kleinere **Nachteile** dieses Verfahrens können dabei in Kauf genommen werden. So wird durch die enge Orientierung an SNOMED der Algorithmus sehr intensiv an diese Nomenklatur und ihre Struktur geknüpft. Die Einteilung von SNOMED in Achsen wurde hier aber nur der Einfachheit halber zur Einteilung der Codes in drei Kategorien großteils übernommen. Prinzipiell kann jedem SNOMED-Code ein eigenes Gewicht 2, 1 oder 0 zugeordnet werden je nachdem es sich dabei um einen Krankheitsbegriff, eine topographische Lokalisierung oder einen der restlichen Bezeichner handelt; die Ermittlung der Reihenfolge erfolgt dann entlang dieses semantischen Gewichtes. Bei Verwendung einer anderen medizinischen Nomenklatur oder einer neueren SNOMED-Version zur Indexierung, bei der Code und Bedeutung nicht mehr in einer solchen Weise korrelieren, würde dies allerdings ein Durchgehen der Einträge im Einzelnen erfordern, um sie den drei semantischen Kategorien zuzuordnen.

Auch kommt es vor, dass die Rechenzeiten durch die getroffene Anordnung bei bestimmten Diagnosen leicht ansteigen. So ist z.B. ein Prostatakarzinom sehr viel schneller zu finden, wenn man alle möglichen Krankheiten der Prostata durchsucht, als wenn alle Lokalisationen eines malignen Tumors überprüft werden müssen. Dieser Zeitverlust spielt allerdings in der Praxis eine sehr untergeordnete Rolle. – Dass durch die Gewichtung der Codes auch in Einzelfällen Schlüsselbegriffe an letzter Stelle für die Suche plaziert werden können, kommt

ebenfalls vor, doch sind diese Fehlgewichtungen nicht systematisch und daher für einen Klassifikationsalgorithmus, der sich an Wahrscheinlichkeiten orientieren muss, nicht zu erfassen. Ausgenommen davon sind Begriffe der L-Achse, die für eine korrekte Gewichtung jeweils einer der drei Kategorien Krankheit, Topographie und Restliches zugeordnet werden müssten, anstatt sie als „Lückenbüßer“ zu verwenden. Da aktuell in der Wissensbasis allerdings 22369 Begriffe der L-Achse zugeordnet werden, die in der Praxis bei Verschlüsselungen nur in Einzelfällen vorkommen und dann meist sehr untergeordneter Bedeutung sind, wurde auf diese aufwendige Bewertung bisher verzichtet.

Einzelne **Verbesserungsmöglichkeiten** für die Gewichtung bzw. Anordnung der SNOMED-Terme wären möglich. Zunächst könnte eine Überprüfung der Einträge von Hand noch einzelne Fehlbeurteilungen ausfindig machen, da bislang die Einteilung entlang der SNOMED-Achsen und innerhalb der Achsen der Einfachheit halber anhand ihrer Reihenfolge erfolgt, die diese Bewertung in vielen Fällen zwar ermöglicht, eigentlich dafür aber nicht vorgesehen ist. – Auch eine mehrstufigere Unterteilung der Gewichte wäre denkbar. So sind etwa Prozedurbegriffe in den Zusatzklassifikationen im ICD-9 (V01.0-V82.9 „Zusatzklassifikation für Faktoren, die den Gesundheitszustand und die Inanspruchnahme von Einrichtungen des Gesundheitswesens beeinflussen“) bzw. im Kapitel XXI des ICD-10 („Faktoren, die den Gesundheitszustand beeinflussen und zur Inanspruchnahme des Gesundheitswesens führen“) sehr viel relevanter, als ihr Rang in der Suchreihenfolge – letzte Stelle – dies ausdrückt.

Des Weiteren wäre auch eine Hierarchie einzelner Erkrankungen zum Zwecke der Unterscheidung von Haupt- und Nebendiagnosen denkbar, da etwa in Arztbrief- und Entlassungsdiagnosen oft zusätzliche Informationen und damit untergeordnete Krankheitsbezeichner vorkommen. Eine Leukämie beispielsweise ist sicher für den Patienten und für die Klassifikation der Hauptdiagnose relevanter als ein begleitender Infekt. Bisher ist die Reihenfolge der SNOMED-Codes innerhalb der drei großen Kategorien Erkrankung, Topographie und Rest aber arbiträr und resultiert in nicht explizit vorgeschriebener Weise aus der Reihenfolge der Wörter innerhalb des Dokuments. Da aber jeder SNOMED-Code, nach dem anhand des beschriebenen Rankings zuerst gesucht wird, den Suchraum für die weiteren Schlüssel festlegt und erheblich reduziert, kann eine Permutation der Wörter im Dokument im Resultat zu einer anderen ICD-Klasse führen. Dies könnte zwar sinnvoll sein, wenn durch eine Neuordnung der Wörter die Bedeutung des Dokumenttextes verändert wird; doch ist die Art, wie der Algorithmus zu einem anderen Ergebnis kommt, nicht logisch begründet, da die Umordnung der SNOMED-Codes auf eine solche etwa veränderte Bedeutung ja keinen Bezug nimmt.

Insgesamt jedoch wird durch die bisherige einfache Gewichtung der SNOMED-Einträge innerhalb des indexierten Dokuments eine sehr praktikable und in den meisten Fällen auch sehr sinnvolle Anordnung für das weitere Retrieval festgelegt.

4.4 Der Retrieval-Prozess

Nachdem nun die notwendigen Vorbereitungen getroffen worden sind, ist das Prinzip des eigentlichen Retrievalvorgangs relativ einfach. Zunächst existiert zu jedem SNOMED-Code (immer inklusive der L-Achse) eine Liste der Identifier sämtlicher Diagnosen des Lerndatensatzes (ICD-Systematik, alphabetisches Verzeichnis und Ergänzungen), deren SNOMED-Indexierung diesen Code mindestens einmal enthalten; ein Auszug aus dieser Liste ist der Tabelle **(12) Clues_ICD-9, Clues_ICD-10, Clues_OPS-301** zu entnehmen. Jeder SNOMED-Code wird also auf eine Menge von ICD-Klassen abgebildet. Diese Mengen werden nun sukzessive in der obigen Reihenfolge nach bestimmten Regeln (4.5) geschnitten. Anschließend (4.6) wird für die verbleibende Menge an ICD-Klassen die Distanz zum ursprünglichen Dokument ermittelt und die bezüglich der noch zu definierenden Metrik am nächsten liegende Klasse als Ergebnis des Retrievals definiert.

Um generell den einem Anfragedokument nächsten Eintrag in einem Suchraums zu ermitteln, ist die einfachste und korrekte Lösung, den Abstand der indexierten Suchanfrage zu sämtlichen Elementen der Dokumentenkollektion zu errechnen, um von allen Alternativen ein Element mit der im Sinne der Metrik minimalen Entfernung ermitteln zu können. Die Reduktion des Suchraums auf immer kleinere Schnittmengen gewährleistet dies indes nicht in jedem Fall. Die Beschränkung auf eine mit hoher Wahrscheinlichkeit relevante Teilmenge des ICD hat aber den Vorzug, den Algorithmus bei vergleichsweise geringem Verlust an Retrievalqualität immens zu beschleunigen und ein Auffinden einer ICD-Klasse in Echtzeit erst zu ermöglichen. Der Konflikt zwischen praktischer Durchführbarkeit des Retrievals und hoher Qualität des Retrievalergebnisses hat so auch an dieser Stelle der Entwicklung eines neuen Lösungsansatzes gedient, der das Optimieren beider Aufgaben im Verbund sehr weitgehend gewährleistet.

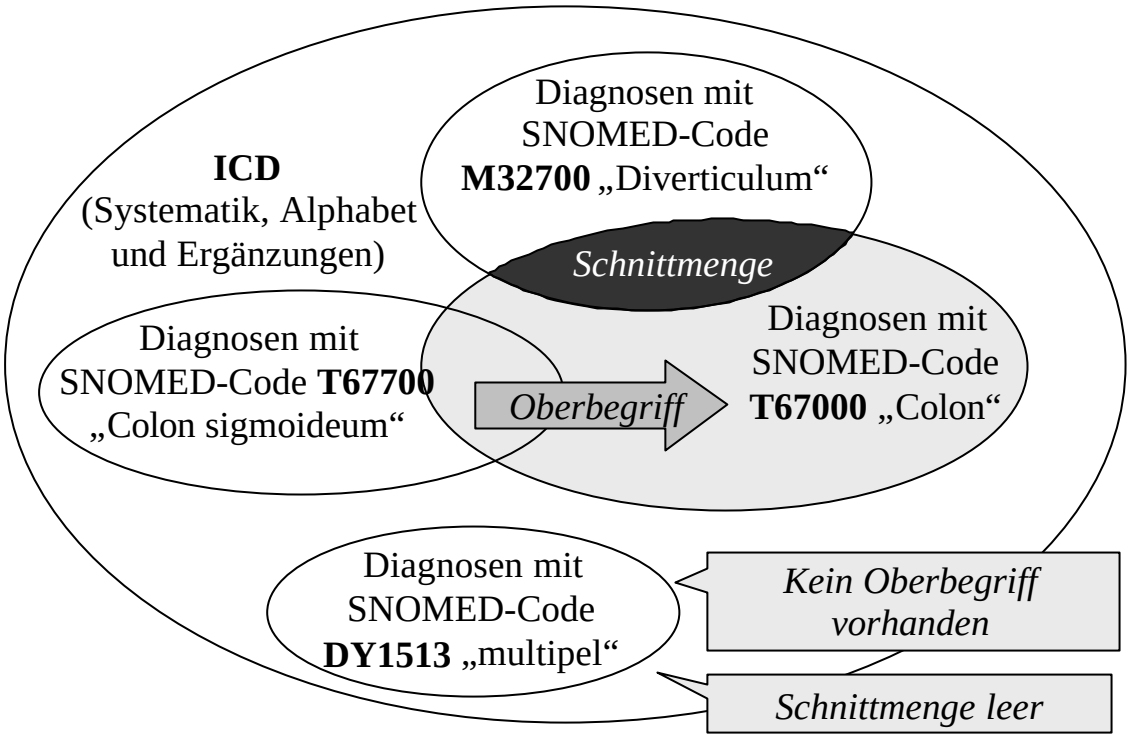
Da das Bilden von Schnittmengen ein kommutativer Vorgang ist, wäre die Reihenfolge des Vorgehens prinzipiell gleichgültig. Das Problem des naiven Ansatzes ist jedoch, dass sich in

vielen Fällen eine nichtleere Schnittmenge, d.h. eine Menge an Diagnosen des ICD-Lerndatensatzes, die die Elemente des SNOMED-indexierten Suchdokuments vollständig abdecken, gar nicht finden lässt, da in der Praxis eine Diagnosenbezeichnung regelmäßig für die Klassifikation überflüssige Informationen enthalten kann und wird. Daher erfolgt das Schneiden der Mengen nach bestimmten Kriterien. – Der wichtigste Gedanke ist dabei, dass eine leere Schnittmenge immer das schlechteste Retrieval-Ergebnis liefern wird, nämlich keines, und daher zu vermeiden ist. Konkret heißt das, dass immer nur dann die Schnittmenge gebildet wird, wenn der Schnitt auch wirklich nichttrivial ist; ansonsten wird die vorige Schnittmenge zunächst belassen. Dadurch ist der Vorgang nicht mehr kommutativ: Die in der Reihenfolge des indexierten Dokuments später auftretenden SNOMED-Schlüssel können nur dann den Suchraum noch eingrenzen, wenn sie im durch die vorigen Schlüssel schon eingegrenzten Suchraum überhaupt noch vorkommen. Begonnen wird der Prozess der Schnittmengenbildung mit dem ganzen Suchraum, also dem ICD-Lerndatensatz aus Systematik, Alphabet und Ergänzungsliste.

Da bei einem solchen rigorosen Vorgehen jeder leere Schnitt das Verwerfen eines ganzen SNOMED-Codes und der durch ihn repräsentierten Information im Dokument nach sich ziehen würde, geht auf diese Art allerdings eine Fülle an Wissen verloren. Dies ist insbesondere dann von großer Relevanz, wenn der zu verwerfende Code deshalb nicht in einer Diagnose des Lerndatensatzes vorkommt, weil die ICD-Klasse, der das Dokument zuzuordnen ist, einen Oberbegriff des SNOMED-Schlüssels enthält. Da das Prinzip jeder Klassifikation ist, viele Einzelfälle unter einem Oberbegriff zu subsummieren, ist diese Konstellation aber sehr frequent. Daher ist es nicht sinnvoll, einen SNOMED-Code völlig zu verwerfen, weil er in der verbleibenden Teilmenge des Suchraums nicht mehr erscheint. Vielmehr muss versucht werden, einen Oberbegriff für ihn ausfindig zu machen. Dieser Gedanke hat dazu geführt, auf den Einträgen von SNOMED II, auf denen das Retrieval ja basiert, eine Hierarchie einzuführen (4.5). Mittels dieser Hierarchie ist es möglich, einen SNOMED-Code dann, wenn er in der durch höher gerankte Codes festgelegten Teilmenge des Suchraums nicht mehr in Erscheinung tritt, durch einen SNOMED-Schlüssel zu ersetzen, der einem Oberbegriff zugehörig ist. In vielen Fällen taucht dann der Code dieses Oberbegriffs in der vorliegenden Teilmenge auf, so dass er zum Bilden der nächsten nichtleeren Schnittmenge verwendet werden kann. Ist dies nicht der Fall, kann der Prozess des Zuordnens eines Oberbegriffs noch ein zweites Mal stattfinden, bevor der Code, da eine weitere Verallgemeinerung dann zu zu vagen Bezeichnungen führen würde, tatsächlich verworfen werden muss. Das Flussdiagramm 4.2 illustriert diesen Teil des Retrievals.

Schließlich verbleibt nach der letzten Schnittbildung eine nichtleere Menge von Elementen des Suchraums, die dem Suchdokument in einigen, nicht aber unbedingt maximal vielen medizinisch relevanten Begriffen oder Oberbegriffen entsprechen. (Die Maximalität wäre wie erläutert nur erreichbar, wenn sämtliche möglichen Permutationen der SNOMED-Codes des indexierten Suchdokuments nach obiger Prozedur untersucht werden könnten.) Auf dieser Menge, die allerdings häufig aus nur einem Element besteht, muss im Folgenden (Abschnitt 4.6) eine weitere Disambiguierung stattfinden, um die dem Dokument nächste ICD-Klasse zu ermitteln. Die übrigen Ergebnisse werden auch hier verworfen. (Da im gesamten Algorithmus zweitrangige Zwischenresultate nicht weiterbehandelt wurden, wäre auch nicht zu erwarten, dass Ergebnisse, die an zweiter, dritter usw. Position stehen, tatsächlich den zweitbesten, drittbesten usw. Resultaten im Sinne unseres Verfahrens entsprächen.)

Wie nach dem beschriebenen Retrieval-Algorithmus mit dem Dokument „Multiple Sigmadivertikel“ aus Beispiel 4.3 verfahren wird, erläutert die Skizze 4.4, die den doch sehr abstrakten Ansatz anhand eines Beispiels verdeutlichen soll. Der SNOMED-Code „**M32700** Diverticulum“ legt zunächst eine Teilmenge des Suchraums fest, auf die sich das weitere Retrieval beschränkt. Da die Teilmenge an ICD-Klassen, die den Begriff „**T67700** Colon sigmoideum“ enthalten, mit dieser Menge disjunkt ist, muss für ihn ein Oberbegriff gefunden werden, in diesem Fall „**T67000** Colon“, so dass der Schnitt nichttrivial wird. Gerade topographische Begriffe müssen oft verallgemeinert werden, da die Unterteilung von Diagnosen nach ihrer topographischen Lokalisation im ICD, wo es hauptsächlich um die Erkrankung als solche geht, oft recht unscharf ist. – Der Code „**DY1513** multipel“, obwohl in der „Disease“-Achse von SNOMED der Kategorie „Rest“ zuzuordnen und daher in der Suchreihenfolge zuletzt, bildet mit der verbleibenden Menge ebenfalls einen leeren Schnitt, kann aber nicht zu einem Überbegriff erweitert werden und wird daher verworfen. Das Retrievalergebnis ist also in der Menge aller ICD-Klassen zu suchen, bei denen es um Divertikel („**M32700** Diverticulum“) des Dickdarms („**T67000** Colon“) geht.



Skizze 4.4: Der Retrieval-Prozess anhand eines Beispiels

4.5 Klassifikationsschritt: Zuordnen eines Oberbegriffs

Die Essenz einer Klassifikation ist das Zusammenfassen von Bezeichnungen zu semantischen Klassen und das Verallgemeinern spezieller Begriffe zu Oberbegriffen. Dieser Mechanismus spiegelt sich, wie Skizze 4.4 illustriert, auch im MedSearch-Algorithmus wieder. Da das MedSearch-Retrieval auf der vorangehenden SNOMED-Indexierung basiert, muss deshalb innerhalb der SNOMED eine Hierarchie definiert werden, die einer Begriff-Oberbegriff-Relation entspricht und eine Baumstruktur implementiert, bei der die Wurzel einen virtuellen Oberbegriff für alle SNOMED-Codes darstellt. Eine solche Struktur ist in vielen Fällen der Architektur von SNOMED II bereits innewohnend, sie wurde allerdings ähnlich wie die Aufteilung in die Kategorien Erkrankung, Topographie und Rest bei der Konzeption von SNOMED international nicht explizit für eine solche Verwendung deklariert und in vielen Fällen auch gar nicht konsequent eingehalten. Wie bereits bei der Definition obiger semantischer Kategorien kann aber auch hier SNOMED dazu dienen, das Erstellen einer solchen Baumstruktur zu erleichtern, indem diese mit einigen Abstrichen durch große Teile der SNOMED innewohnenden Hierarchie aufgefüllt wird. Der Aufwand für die Analyse und Ermittlung einer hierarchischen Begriff-Oberbegriff-Struktur wäre sonst beträchtlich. – Für neuere Versio-

nen von SNOMED wie SNOMED RT und SNOMED CT, die in deutscher Sprache allerdings noch nicht erhältlich sind, existiert explizit eine solche Hierarchie, deren Verwendung für das effiziente Codieren medizinischer Konzepte vielversprechend ist [Bousquet 00]. Ein interessanter Ansatz, um ein semantisches Lexikon aus anderen Quellen wie UMLS zu konstruieren und für Algorithmen zur Analyse medizinischer Texte zu nutzen, findet sich in [Johnson 99].

In Einzelfällen wäre eine korrekte Hierarchiebildung auch nicht unidirektional, wie in einer Baumstruktur, sondern müsste den Schluss von einem Begriff auf verschiedene Oberbegriffe zulassen, die nicht notwendigerweise untereinander durch eine Begriff-Oberbegriff-Relation verknüpft sind. Ist der Oberbegriff für „**D67040** akute Glomerulonephritis“ tatsächlich die sich an den SNOMED-Identifiern orientierende, topographisch motivierte Verallgemeinerung „**D67000** glomeruläre Krankheit“, oder nicht vielmehr der den Charakter der Krankheit beschreibende Begriff „**M40000** Entzündung“? Da aber auch in den meisten solchen Fällen relativ gut festgelegt werden kann, welcher mögliche Oberbegriff im Hinblick auf eine Klassifikation von Erkrankungen der sinnvollste ist, wurde auf die Implementierung eines solchen multidirektionalen Netzwerkes zugunsten der Vereinfachung des Algorithmus verzichtet. (Einen möglichen Ansatz für relationale semantische Netzwerke auf medizinischen Ontologien bietet unter anderem [Schulz 98a].) Wie an vielen Stellen innerhalb des Algorithmus stellt eine weitere Verfeinerung auch hier ein Analogon zur Erhöhung seiner Sensitivität dar (mehr potentielle Resultate werden erfasst), die mit einer gewissen Verringerung seiner Spezifität einhergeht (mehr irrelevante Möglichkeiten werden betrachtet) und nur dann einen echten Gewinn bringen kann, wenn die Anzahl der Fälle, die von der Verfeinerung betroffen sind, groß genug ist; ein heuristisches Argument, das vielen Entscheidungen bei der Entwicklung des Retrievals zu Grunde lag und sie motiviert, wenn der Aufwand für das Austesten der einzelnen Alternativen zu groß erschien.

Die erwähnte der SNOMED II zu weiten Teilen innewohnende Begriffshierarchie, die für die Generierung einer hierarchischen Baumstruktur genutzt wurde, ließ sich bereits in obigem Beispiel erkennen. Sie beruht auf der Tatsache, dass in SNOMED international oft Unterbegriffe, die einen allgemeineren Begriff verfeinern, eine bis auf die letzte Stelle, die ungleich Null ist, identische Notation besitzen. So wäre beispielsweise für den Begriff „**D75260** Vogelzüchter-Lunge“ der Eintrag „**D75200** Docker-Lunge“ ein Oberbegriff, und für diesen wiederum der SNOMED-Eintrag „**D75000** Krankheiten des respiratorischen Systems“. Die nächste Stufe hingegen, „**D70000** Krankheiten des kardiovaskulären Systems“, ist keine Verallgemeinerung und liefert einen falschen Oberbegriff. Eine Hierarchie nach diesem Muster ist

ganz allgemein in weiten Teilen der SNOMED II umgesetzt worden, liegt aber offensichtlich nicht in jedem Einzelfall vor. Genauso etwa wäre für „**E49220** Vogel“ der Eintrag „**E49200** Chordata (Wirbeltiere)“ noch eine Verallgemeinerung; der nächste Eintrag **E49000** hingegen fehlt, und was dann folgt ist „**E40000** Fungus“, also „Pilz“, sicherlich kein Oberbegriff für Vögel. So sollten ganz allgemein fehlende Einträge oder „Blanks“ wie in diesem Fall **E49000** bei der Bildung einer hierarchischen Struktur nicht übersprungen werden.

Da der Aufwand für das Erstellen einer eigenen Hierarchie zu groß gewesen wäre und die Suche nach einem Oberbegriff in den meisten Fällen von deutlichem Vorteil ist, da derzeit auf anderem Wege die vorhandene Information nicht interpretiert werden könnte, wurde eine Hierarchie in Anlehnung an die Struktur von SNOMED II definiert. Dabei musste akzeptiert werden, dass es in einzelnen Fällen (wie im ersten Beispiel) zu falschen Oberbegriffen kommen kann, wenn das Rezept zum Uminterpretieren der SNOMED-Codes zum Zwecke des Aufbaus einer Hierarchie fehlschlägt.

Die Definition für diesen Prozess ist also folgende: Um ausgehend von einer SNOMED-Notation zur Notation eines Oberbegriffs zu gelangen, muss die letzte Stelle des SNOMED-Identifiers, die ungleich Null ist, durch eine Null ersetzt werden. Entsteht dadurch ein gültiger SNOMED-Code, so entspricht dieser dem Oberbegriff; wenn nicht, kann kein Oberbegriff ermittelt werden (bzw. nur die virtuelle Wurzel des Baumes wäre Oberbegriff, aber damit lässt sich nichts anfangen).

Ausgenommen von dieser Definition sind einige Teile der SNOMED, die zumeist am „Ende“ der Achsen auftauchen und eine Art Sammelbecken für Begriffe bilden, die in einer umfassenden medizinischen Begriffsammlung aufgelistet sein müssen, aber oft nur im ferneren Sinne einer der Achsen zuzuordnen sind. Da sie in den meisten Fällen weder Erkrankungen, Prozeduren oder topographische Lokalisierungen beschreiben, sind diese Einträge im Ranking des indexierten Suchdokuments ohnehin an nachrangiger Stelle zu finden. Im Einzelnen ausgeschlossen werden mussten:

Von der D-Achse: Alles ab DY1000 (inklusive).
Von der E-Achse: Alles ab EY0000 (inklusive).
Von der F-Achse: Alles ab FYX000 (inklusive).
Von der J-Achse: Nichts.

Von der L-Achse: Alles (die zusätzliche L-Achse ist in keiner Weise hierarchisch strukturiert).

Von der M-Achse: Alles ab MX0000 (inklusive). Die Werte ab M80000 haben einen Sonderstatus (siehe unten).

Von der P-Achse: Nichts.

Von der T-Achse: Alles zwischen T00200 (inklusive) und T00400 (exklusive).

Einen Sonderstatus besitzt hierbei die Morphologie der Neubildungen. Die Notationen M80000 (inklusive) bis MX0000 (exklusive) bestehen aus dem Buchstaben M und einer fünfstelligen duodekadischen Ziffernfolge, von denen die ersten vier Stellen die Histologie der Neubildung kennzeichnen und die fünfte Stelle ihren Charakter. So bezeichnet etwa eine 0 an fünfter Stelle eine benigne Neubildung, eine 2 ein Carcinoma in situ, eine 3 ein malignes Geschehen (primärer Sitz), eine 6 eine Metastase usw. Diese letzte Stelle muss in jedem Fall erhalten bleiben, da eine gutartige Neubildung (z.B. **M81610** Gallengangszystadenom) kein Sammelbegriff für alle Arten von Neubildungen der betreffenden histologischen Gattung darstellt (z.B. **M81613** Gallengangszystadenokarzinom). Stattdessen muss die letzte der vorderen vier Ziffern, die ungleich Null ist, durch eine Null ersetzt werden, um – zumindest in vielen Fällen – zu einem Oberbegriff zu gelangen (z.B. **M81603** Gallengangskarzinom). –

Es ist zu betonen, dass nicht alle Begriff-Oberbegriff-Relationen von SNOMED-Einträgen, die existieren würden, durch das so definierte Verfahren auch entdeckt werden. Beispiel wäre etwa „**F85830** Schlafstörung“ als Oberbegriff von „**F85920** Durchschlafstörung“, oder „**F70700** Hypertonie“ als Oberbegriff von „**D73800** essentielle Hypertonie“. Abhilfe schaffen würde hier das aufwendige manuelle Erstellen eines Thesaurus, welcher wichtige Begriff-Oberbegriff-Relationen enthält. Eine intelligente, aber leider ebenfalls recht fehlerbehaftete und nicht unaufwendige Alternative wäre das maschinelle Codieren von SNOMED durch „sich selbst“, d.h. das Verschlüsseln jeden SNOMED-Eintrags, soweit möglich, durch andere SNOMED-Einträge (eine ähnliche Strategie wird von SNOMED RT verfolgt [Spackman 98]). So könnte man etwa beim Indexieren von „**F70900** Renale Hypertonie“ durch „**T71000** Niere“ und „**F70700** Hypertonie“ zu zwei Oberbegriffen gelangen, von denen ebenfalls automatisch festgelegt werden kann, dass der zweite (da in der Kategorie „Erkrankung“ gelegen) der für die Klassifikation von Diagnosen ausschlaggebende ist.

Ein weiteres Problem des jetzigen Vorgehens ist sicher, dass beim Auffinden auch nur einer Diagnose, die den gesuchten SNOMED-Code enthält, alle weiteren, die eventuell einen Ober-

begriff enthalten, unter den Tisch fallen. Andererseits ist eine extensivere Verwendung von Oberbegriffen von einem zunehmenden Präzisionsverlust begleitet.

4.6 Metrik des Suchraumes

Für die sich nach 4.4 und 4.5 ergebenden SNOMED-codierten ICD-Lerndaten muss nun bestimmt werden, welche Distanz sie im Einzelnen zum indexierten Suchdokument haben. Dazu wird auf dem Suchraum des MedSearch-Retrieval, dem indexierten ICD-Lerndatensatz, eine „gerichtete“ Metrik eingeführt. Diese ist wie folgt definiert:

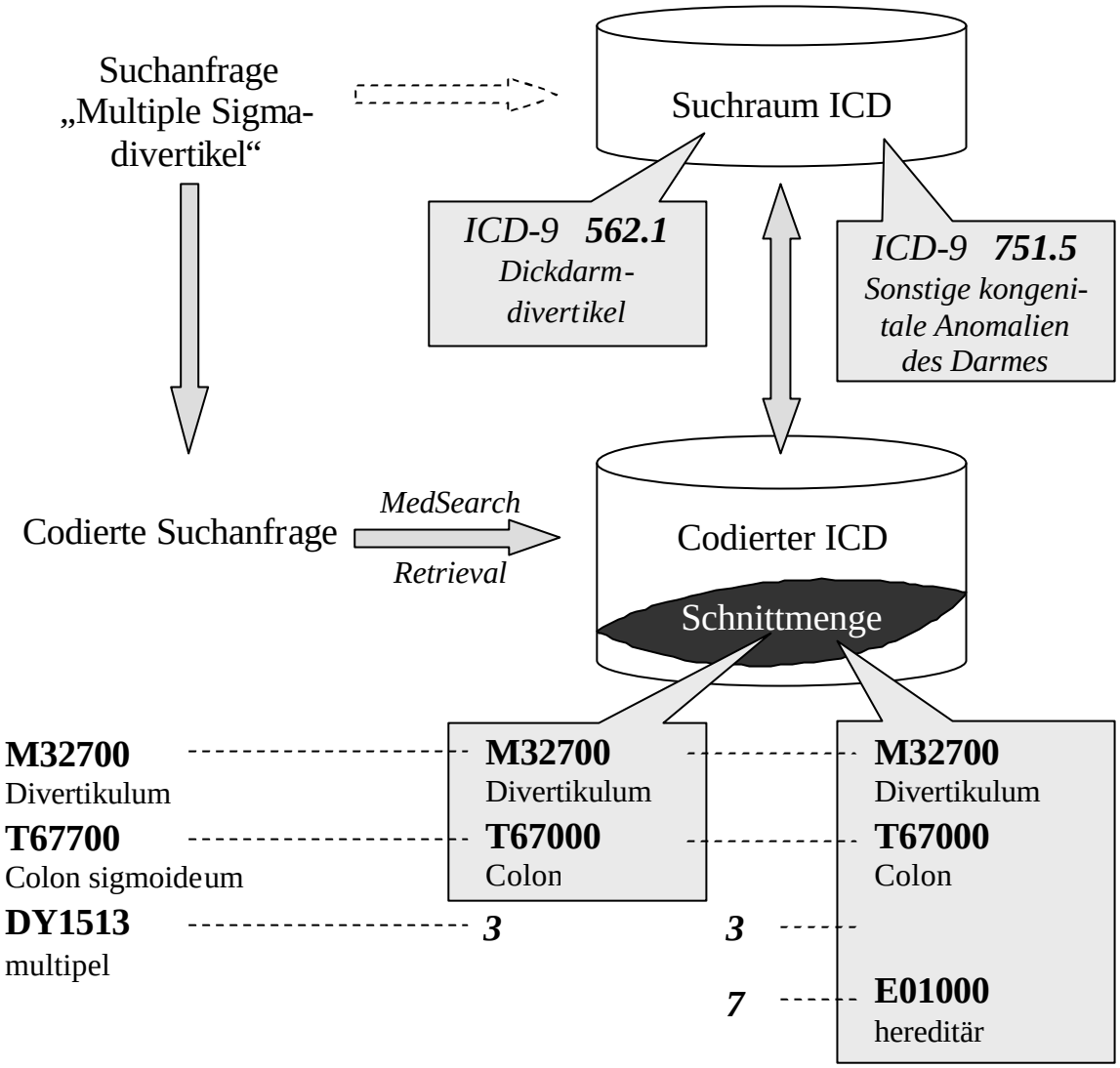
- Ist ein indexiertes Suchdokument mit einem Element des Suchraumes (d.h. einer Folge von SNOMED-Notationen) bis auf die Reihenfolge der Notationen identisch, so ist ihre Distanz 0.
- Für jedes Vorkommen eines SNOMED-Eintrags im indexierten Suchdokument der nicht, auch nicht in Form eines Oberbegriffs, im zu vergleichenden Element des Suchraumes vorkommt, steigt die Distanz (beginnend bei 0) um 3.
- Für jedes Vorkommen eines SNOMED-Codes im zu vergleichenden Element des Suchraumes der nicht, auch nicht in Form eines Unterbegriffs im Sinne der Begriff-Oberbegriff-Hierarchie, im indexierten Suchdokument vorkommt, steigt die Distanz um 7. Das Ergebnis wird als Abstand des Elements des Suchraumes vom indexierten Suchdokument bezeichnet.

Da die Suchanfrage sehr häufig zusätzliche, für die Klassifikation irrelevante Informationen enthält, die in den Diagnosen des ICD-Lerndatensatzes enthalten Information hingegen möglichst vollständig abgedeckt werden sollte, werden als Entfernungsmaß die beiden unterschiedlichen Summanden 3 und 7 verwendet. Über die Höhe dieser Gewichtung kann die Distanz zu bestimmten Dokumenten reguliert werden. Die Wahl der Gewichte 3 und 7 im Algorithmus hat zur Folge, dass ein indexiertes Suchdokument mit zwei SNOMED-Notationen, die im zu vergleichenden Element des Suchraums nicht vorkommen und das also Zusatzinformationen enthält, die in aller Regel für die Klassifikation unbedeutend sind (Distanz 3+3 = 6), diesem Element immer noch näher liegt als einem anderen Element des Suchraums, von dem ein SNOMED-Code in der Indexierung der Suchanfrage fehlt (Distanz 7) und daher ein ganz entscheidender Teil an Information, da sämtliche Diagnosen aus der ICD-Systematik,

dem alphabetischen Verzeichnis und der Ergänzungsliste keine überflüssigen Informationen enthalten sollten.

Da aufgrund dieser unterschiedlichen Gewichtung die für eine echte Metrik notwendige Voraussetzung der Symmetrie verletzt ist, das beschriebene Maß aber dennoch in plausibler Weise eine Distanz zu Elementen des Suchraumes definiert, soll es mit dem Begriff der „gerichteten“ Metrik bezeichnet werden.

Das Suchdokument „Multiple Sigmadivertikel“ soll auch hier wieder als Beispiel dienen (Skizze 4.6). Die korrekte Klasse, in die es fällt, wäre „ICD-9 **562.1** Dickdarmdivertikel“. Allerdings können Divertikel auch angeboren sein; in diesem Falle wäre die korrekte Klasse „ICD-9 **751.5** Sonstige kongenitale Anomalien des Darmes“. Für diese Klasse gibt es im alphabetischen Verzeichnis tatsächlich die Beispieldiagnose „Angeborenes Dickdarmdivertikel“. Beim Vergleich der SNOMED-Indexierung des Suchdokuments mit diesen Diagnosen des Suchraums ergibt sich in beiden Fällen keine völlige Übereinstimmung, da das Suchdokument die Zusatzinformation „multipel“ enthält. Diese wird in beiden Fällen mit dem Gewicht 3 bewertet. Im zweiten Fall enthält die zu vergleichende Diagnose aber zusätzlich den Begriff „angeboren“, was als SNOMED-Eintrag „**E01000** hereditär“ mit einer weiteren 7 in die Gewichtung eingeht, so dass die Distanz des Suchdokuments zur zweiten Diagnose des Lerndatensatzes mit insgesamt 10 gegenüber der Distanz 3 ungleich größer ist.



Skizze 4.6: Berechnung von Distanzen zu einem Dokument

Bei der Definition der Distanz sind nicht nur was die beiden Gewichte betrifft viele andere Abstufungen denkbar. So stößt man sich etwa am Beispiel in Skizze 4.6 an der Tatsache, dass zwischen einem Begriff und seinem Oberbegriff, der sich von ihm selbst bei korrekter Zuordnung (die durch die in 4.5 definierte Hierarchie in etlichen Fällen nicht einmal gewährleistet ist) doch erheblich unterscheiden kann, gar kein Unterschied gemacht zu werden scheint. Eine zusätzliche Distanzmessung diesbezüglich wäre aber nur dann sinnvoll, wenn Diagnosen aus dem Suchraum, die lediglich einen Oberbegriff enthalten, generell mitbeurteilt würden, also auch in Fällen in denen Beispieldiagnosen existieren, in denen der Begriff selbst vorkommt. So wie das Retrieval aber definiert wurde, kommen Oberbegriffe in der Schnittmenge nur

dann vor, wenn sie in jedem einzelnen der Elemente gleichermaßen vorliegen, so dass keine Gewichtung hier die Differenz zwischen den Distanzen beeinflussen wird.

Hingegen könnten wiederum die einzelnen SNOMED-Codes, die in Dokument oder Beispieldiagnose fehlen, nach ihrer Semantik gewichtet werden. Hier würde sich etwa eine Einteilung wie in Abschnitt 4.3 in Krankheitsbegriffe, topographische Bezeichnungen und alles übrige anbieten. So würde nicht nur die Differenz an SNOMED-Einträgen, sondern auch ihre Bedeutung in die Distanzmessung eingehen, da ein Fehlen von Begriffen, die unter „Erkrankungen“ fallen, sicher schwerer wiegen würde als bei topographischen Lokalisationen und so fort. Allerdings wurde genau dieser semantischen Bedeutung schon durch die Anordnung der SNOMED-Codes in hohem Maße Rechnung getragen, so dass eine zusätzliche Betonung dieser Semantik in den häufigen Fällen, in denen sie bedeutsam ist, meist keine zweite Verbesserung bringt, und dafür in den selteneren Fällen, in denen sie kontraproduktiv ist, da im Suchdokument ausnahmsweise ein weniger hoch eingestufter SNOMED-Eintrag entscheidend ist, sich um so mehr negativ auswirken würde.

Welche Verbesserungen an der Bestimmung der Distanz zum Suchraum sinnvoll sind, hängt von der Erfahrung mit dem System MedSearch ab. Wahrscheinlich aber sind feinere Unterschiede nicht von zu großem Belang, da die echten Grenzfälle sehr selten sind. Die wichtigen Entscheidungen fallen bereits in 4.3, 4.4 und 4.5, und die in der letzten Schnittmenge verbleibenden Diagnosen stammen oft genug einer einzigen ICD-Klasse ab oder sind ohne zu penetrante Wortklaubereien (im wörtlichen Sinne) voneinander abzugrenzen.

Da mit der Definition der „gerichteten“ Metrik nun das Fällen des Lots auf den Suchraum möglich ist und der Rückweg vom Suchraum in den ICD bereits in Abschnitt 4.1 definiert wurde, ist hiermit das MedSearch Retrieval abgeschlossen. Vor der Evaluation des Algorithmus, die sich im Weiteren anschließt, folgt nun zunächst ein Kapitel über einige technische Details der Umsetzung von MedSearch, die zur Vereinfachung der Beschreibung von Indexierung und Retrieval in Kapitel 3 und 4 ausgegliedert wurden.

5 Realisation von MedSearch: Technische Hinweise

Die praktische Umsetzung des in den methodischen Kapiteln ausgeführten MedSearch-Algorithmus hängt mit vielen technischen Einzelheiten zusammen, die für seine Definition ausgeklammert werden konnten. Da diese Arbeit aber als Anregung zur weiteren Forschung dienen soll in einem Bereich, an dem das Interesse und in dem die Umsetzungsmöglichkeiten stetig steigen, werden im Folgenden einige technische Aspekte und Erfahrungen beleuchtet, die für eine Realisierung des Verfahrens von Nutzen sind.

Eng verknüpft sind dabei in der Praxis die Möglichkeiten eines solchen Ansatzes mit den Gegebenheiten des Systems. An vielen Stellen in Kapitel 3 und 4 wurde darauf hingewiesen, dass bestimmte Varianten nur deshalb ausgeklammert wurden, weil das Ziel einer automatischen Klassifikation in Echtzeit (damit sollen Zeiträume bis zu 1 s bezeichnet werden) andernfalls aufgrund begrenzter Systemleistung nicht zu erreichen wäre. Besonders der Ansatz der Schnittmengenbildung im Retrievalteil wurde sehr auf die rasche Eingrenzung des Suchraumes und damit das Optimieren der Suchzeiten zugeschnitten. Einbußen an der Retrievalqualität sind bei solchen Beschränkungen im Allgemeinen die Folge. Mit der jetzigen Implementation von MedSearch, die in ihren Ursprüngen noch auf einem 486er Prozessors (33 MHz) entwickelt wurde, sind Echtzeiten bereits ab der Verwendung eines Pentium 2 (133 MHz) möglich; inzwischen sind wesentlich leistungsfähigere Systeme im klinischen Routineeinsatz. Von technischer Seite besteht daher für Weiterentwicklungen ein großer Spielraum. –

Um in der Praxis zum Einsatz kommen zu können, wurde MedSearch nach außen hin mit Schnittstellen umgeben, die bestimmte Anfragen ermöglichen und Resultate liefern. Diesen äußeren Rahmen beschreibt Abschnitt 5.1. Nach innen hin wurde, da eine Überwachung der Einzelschritte des Algorithmus und eine Fehlerkontrolle für die Evaluation und Weiterentwicklung möglich sein sollte, eine Protokollierung der Arbeitsschritte umgesetzt (Abschnitt 5.2). Da MedSearch als morphemorientierter, lexikabasierter Klassifikationsalgorithmus sein Hintergrundwissen aus einer ganzen Reihe von Lexika bezieht, sind zu seiner Beurteilung Kenntnisse über den zu Grunde liegenden Lerndatensatz erforderlich, der in Abschnitt 5.3 näher definiert wird. Einige Hintergrundinformationen zur Struktur und möglichen Implementierung des Morphemlexikons werden in 5.4 dargelegt, bevor Abschnitt 5.5 dieses Kapitel mit einer Studie über das Wachstumsverhalten von Morphemlexika abschließt.

5.1 Die Architektur des Systems MedSearch

Das System MedSearch besitzt nach außen hin eine einzige Schnittstelle, auf die sowohl Anwender als auch andere Applikationen zugreifen können. Über sie werden Suchdokumente übergeben sowie Anweisungen zum Ermitteln der nächstgelegenen ICD- oder OPS-301-Klasse. Auch Teilergebnisse wie die Vorverarbeitung, morphologische Analyse oder SNOMED-Indexierung des Dokuments können angefordert werden, was für Anwendungen mit anderen Zielsetzungen und insbesondere für das Testen von Zwischenstufen des Algorithmus und den Vergleich mit anderen Retrievalverfahren von Nutzen ist (vgl. Kapitel 6).

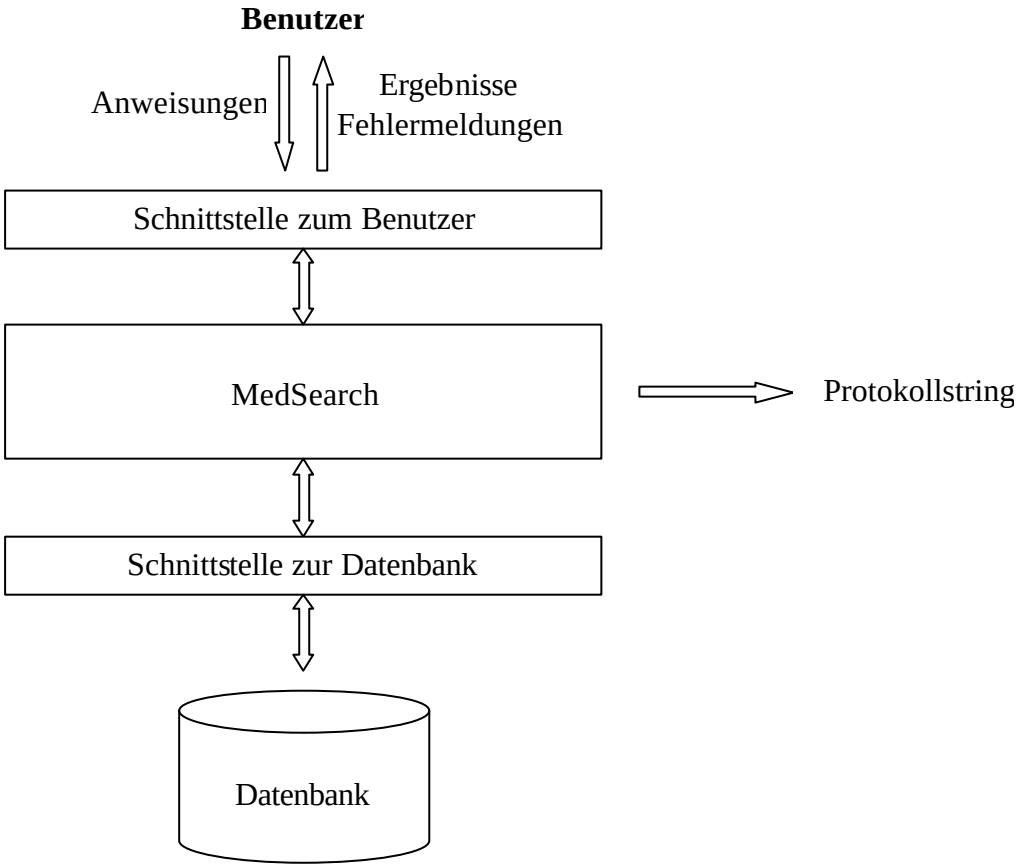
Auch die Verwaltung der Datenbasis sollte im Prinzip ausschließlich über diese Schnittstelle erfolgen. Dazu können dem System neue Daten für die zu Grunde liegenden Lexika übergeben werden oder Anweisungen, um Daten zu lesen oder zu entfernen. Sinn der Verwaltung der Datenbasis durch das System selbst ist eine mit den Systemanforderungen kongruente Speicherung der Informationen, da bei Änderungen in einem Datensatz in aller Regel weitere Metadaten wie Zeiger sowie die vorverarbeiteten, morphologisch analysierten oder SNOMED-indexierten Stammdaten mit beeinflusst werden, die MedSearch automatisch generieren und entfernen kann.

Diese Schnittstelle ist als eigenständiges Modul implementiert, das Anweisungen des Anwenders in interne Befehle übersetzt und im Anschluss an ihre Ausführung eine resultierende Zeichenkette zurückgibt, die das Ergebnis der Klassifikation oder anderweitigen Anordnung enthält, sowie eine Fehlervariable, die auf Probleme aufmerksam macht, wie sie vor allem bei Operationen an der Datenbasis auftreten können.

Intern existiert eine zweite Schnittstelle zur Datenbasis hin, die ebenfalls von einem eigenständigen Modul gebildet wird. Das System MedSearch selber übergibt an dieses Modul Befehle, um Daten zu lesen oder zu schreiben, und das Modul übersetzt diese Befehle zur Datenbasis hin. So ist ein Austausch der Datenbasis theoretisch möglich, indem man die Schnittstelle neu programmiert. Allerdings sind die Befehle, die die Schnittstelle weiterleiten kann, ziemlich genau auf die Möglichkeiten von ACCESS, der von MedSearch verwendeten relationalen Datenbank, zugeschnitten (so etwa die Möglichkeit, vorindexierte Daten rasch durchsuchen zu können). Darüber hinaus sind Teile der Wissensbasis, so etwa die Regeln über die anhand SNOMED gewonnene Einteilung von Begriffen in semantische Kategorien und hierarchische Strukturen, explizit in den Algorithmus mit eingeflossen, die prinzipiell durch ei-

gene Lexika repräsentiert und ausgegliedert werden müssten (da die Regeln zur Berechnung dieser Kategorien und Hierarchien anhand der SNOMED-Notation sehr einfach ist, wurde eine Aufnahme in den Algorithmus wegen der deutlich längeren Zugriffszeiten bei Zwischenschaltung eigenständiger Lexika vorgezogen).

Die folgende Skizze 5.1 soll den erläuterten Aufbau von MedSearch in der Sicht des Benutzers auf das System zusammenfassen:



Skizze 5.1: Architektur von MedSearch

5.2 Protokollierung der Verarbeitungsschritte

Um die Vorgänge innerhalb von MedSearch nachvollziehen können, führt der Algorithmus über die einzelnen Verarbeitungsschritte Protokoll. Diesem Protokoll werden dabei zunächst Befehlsaufruf (m für *main*: Hauptanweisung) und Parameter (i für *item*) übergeben. Wesentlich ist, dass der Algorithmus nicht auf der Ebene der einzelnen Zeichen ansetzt, sondern als

„Atome“ die Wörter (1…n) des Suchausdrucks betrachtet. Anschließend erfolgt als erster Befehl (d für *do*) die Initialisierung: Der Startparameter „0“ wird durch die Begriffe $i_1 \ldots i_n$ ersetzt (der senkrechte Strich symbolisiert den Operator des Ersetzungsvorgangs); dabei ermöglichen es die Identifier, mehrfach vorkommenden identische Begriffe zu separieren. Anschließend werden die Elemente $i_1 \ldots i_n$ nach den in Kapitel 3 und 4 beschriebenen Schritten des Algorithmus bearbeitet; die sich ergebenden Ausdrücke werden jeweils mit neuen Identifiern im Protokoll gespeichert und das Ersetzen eines Wortes wird wiederum durch einen Ersetzungsbefehl dokumentiert, der die Identifier der jeweiligen Elemente auf die ihrer Bearbeitung projiziert. Zusätzlich werden Kommentare (c = *comment*) über den Ablauf des Algorithmus sowie bei den *do*-Anweisungen über den Grund des Ersetzens hinzugefügt. Des weiteren erfolgt zu Dokumentationszwecken die Protokollierung der für die einzelnen Abschnitte verbrauchten Systemzeit (t für *time*). Im Ergebnis gleicht das Protokoll einem kleinen Programm, anhand dessen die verschiedenen Zwischenstufen der Bearbeitung für eine Fehlersuche rekonstruiert werden können. Dabei ist die Möglichkeit des Aufsplittens in mehrere Varianten mit Angabe von Wahrscheinlichkeiten (<1> steht für 100 %) für künftige Weiterentwicklungen des Algorithmus bereits vorgesehen. – Ein typisches Beispiel eines solchen Arbeitsprotokolls ist in Abbildung 5.2 dargestellt.

Das Protokoll dient einer durchsichtigen Programmgestaltung. Der Entwickler kann die einzelnen Schritte des Algorithmus besser nachvollziehen, und insbesondere lässt sich erkennen, welche ursprünglichen Wörter zu welchen Schlüsseln beigetragen haben.

Weiterer Vorteil ist, dass Zwischenergebnisse gespeichert werden und gegebenenfalls abrufbar sind, falls vom Benutzer verschiedene Resultate wie die SNOMED-Indexierung und ICD-Klassifikation angefordert werden. Mit Hilfe des Protokolls kann jederzeit auf früher berechnete Teilergebnisse zugegriffen werden.

```
m1 : Klassifikation
c1 : Beginn Verschlüsselung
c2 : Beginn Diagnose
i1 : 12/93
i2 : Hinterwandinfarkt
d1 : 0|1 2{Initialisierung}
c3 : Ende Diagnose
t1 : 0 ms
c4 : Beginn Vorverarbeitung
i3 : [12]/[93]
d2 : 1|3{Ganze Zahl}
i4 : alt
d3 : 3|4{Datum}
i5 : hinterwandinfarkt
d4 : 2|5{Zeichen reduzieren}
c5 : Ende Vorverarbeitung
t2 : 1157 ms
c6 : Beginn Morphologie
i6 : alt{521,2,4}
d5 : 4|( 6 ){Zerlegung}
i7 : hinter{3801,2,1}
i8 : wand{94,2,1}
i9 : infarkt{1489,2,1}
d6 : 5|<1> ( 7 8 9 ){Zerlegung}
c7 : Ende Morphologie
t3 : 1157 ms
c8 : Beginn Indexierung
d7 : ( & )|{Klammerung aufheben}
i10: M54750 alter Infarkt
d8 : 6 & 9|10{Indexieren}
i11: T00315 hinter
d9 : 7|11{Indexieren}
i12: T00315 dors. {S: hinter}
d10: 11|12{Vorzugsbezeichnung}
i13: TYX460 Wand
d11: 8|13{Indexieren}
i14: TYX460 Paries {S: Wand}
d12: 13|14{Vorzugsbezeichnung}
c9 : Ende Indexierung
t4 : 2315 ms
c10: Beginn Synonyme
c11: Ende Synonyme
t5 : 2315 ms
c12: Beginn Klassifikation
i15: 412 Alter Myokardinfarkt
d13: 10 12 14|15
c13: Ende Klassifikation
t6 : 2315 ms
c14: Beginn Ergebnis
r1 : 412 Alter Myokardinfarkt
c15: Ende Ergebnis
c16: Ende Verschlüsselung
```

Beispiel 5.2: Typisches Klassifikationsprotokoll

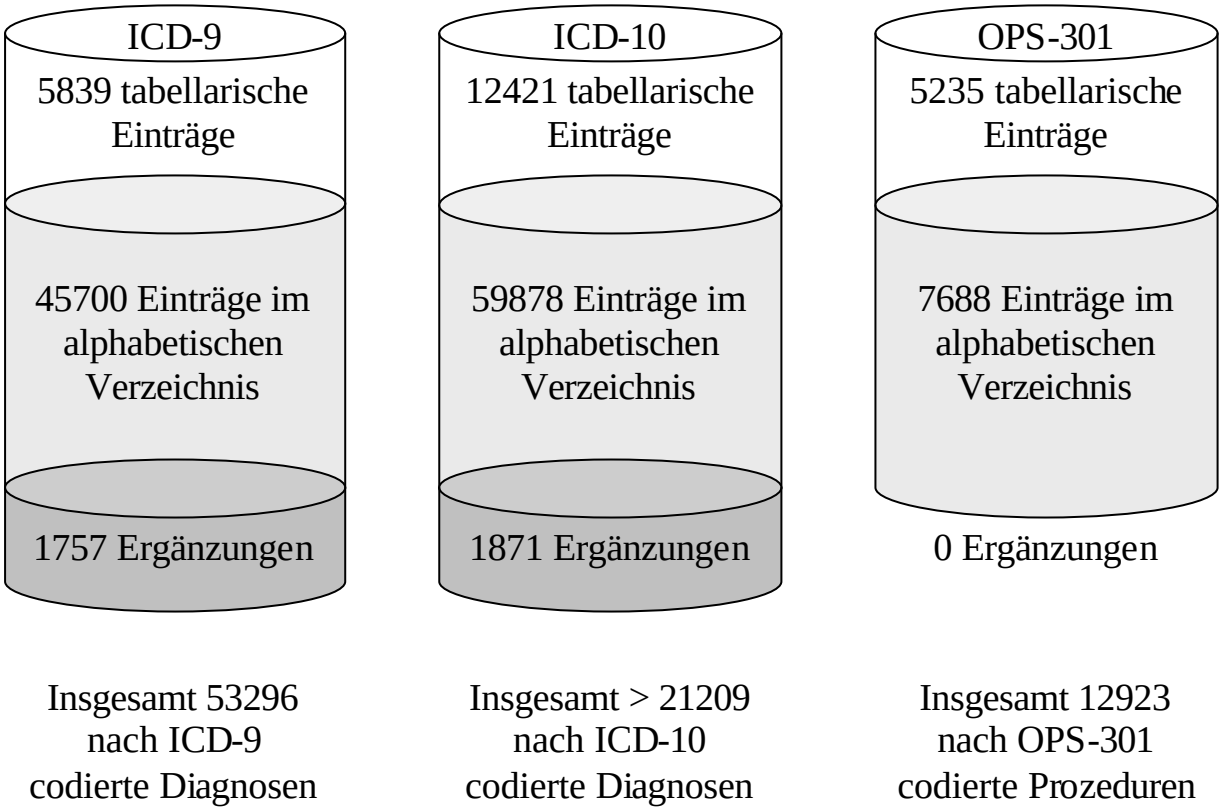
5.3 Der Lerndatensatz von MedSearch

Nachdem in Kapitel 3 und 4 die MedSearch zu Grunde liegenden Datensätze immer wieder erwähnt wurden, die die Basis für den lexikabasierten Ansatz bilden, sollen ihr Inhalt und Umfang im Folgenden kurz vorgestellt werden. Dieser Überblick soll unter anderem dazu dienen, bei der Entscheidung über eine Implementation und Weiterentwicklung des Ansatzes die richtige Einschätzung der Erstellung und Pflege des Datenstammes zu treffen, die im Vorfeld unbedingt bedacht werden sollten.

Zum Lerndatensatz von MedSearch gehören im engeren Sinne die Daten, die den Suchraum des Retrievals definieren. Dies sind zunächst die systematischen Verzeichnisse des ICD-9 und ICD-10 sowie der Prozedurkatalog nach OPS-301, von denen im Falle des ICD alle Diagnosen mit vierstelliger Notation sowie alle Einträge mit drei Stellen, von denen es keine Aufteilung in Viersteller gibt, aufgenommen wurden. Für die ICD-9 sind dies insgesamt 5839, für die ICD-10 12421 Einträge. Vom OPS-301 wurden alle Prozeduren mit fünfstelliger Notation ausgewählt, sowie alle mit einem vierstelligen Klassenbezeichner, der nicht in Fünfsteller unterteilt ist; ihre Anzahl beträgt 5235.

Zusätzlich zu jeder dieser Listen existiert ein alphabetisches Verzeichnis, in dem eine große Anzahl an Diagnosen aufgelistet ist, die beispielhaft für bestimmte Klassen stehen. Dieses Alphabet umfasst im Falle des ICD-9 45700, für den ICD-10 59878 Einträge. Im Falle des OPS-301 (Version 1.1) wurde eine auf der Grundlage der Fassung des Deutschen Ärzte-Verlags erschienene Bearbeitung eines alphabetischen Verzeichnisses verwendet, die 7688 maximal fünfstellig verschlüsselte Datensätze enthält. Um dieses Material verwenden zu können, musste die Einschränkung der OPS-301-Klassifikation auf fünf Stellen der zum Teil sechsstelligen Notation erfolgen; eine Aufnahme der sechsstelligen systematischen Einträge ist daher unterblieben (die Reduktion auf eine fünfstellige Notation wäre alternativ möglich).

Des weiteren existieren zu ICD-9 und ICD-10 lokale Ergänzungslisten mit Diagnosen, die weder in den tabellarischen Listen noch im Alphabet verzeichnet sind, ebenfalls jeweils mit dem korrekten Klassenbezeichner. Für ICD-9 sind dies 1757, für ICD-10 1871 Listeneinträge. Alphabet und Ergänzungsliste bilden zusammen mit der Systematik den jeweiligen Lerndatensatz für das Retrieval, wie er in Abbildung 5.3.1 skizziert ist.



Skizze 5.3.1: Der Lerndatensatz von MedSearch

Keine Anwendung beim Erstellen des Lerndatensatzes fand die Möglichkeit, Teile der ICD-9 und ICD-10-Notationen ineinander zu überführen ([Schulz 98b, Zaiß 96]). Einerseits sind die entsprechenden Diagnosentexte des ICD-9-Alphabets ohnehin in das alphabetische Verzeichnis des ICD-10 eingeflossen; ähnliches gilt für die Ergänzungsliste. Außerdem spiegelt die Anzahl an Einträgen pro ICD-Klasse in grober Weise die Häufigkeit des Krankheitsbefundes wider, was das Auffinden gängiger Diagnosen wahrscheinlicher macht; diese Verteilung würde durch die Hinzunahme von Einträgen in selektiven Bereichen des ICD gestört. Eingesetzt hingegen wurde die Konvertierungstabelle, um einen ersten, allerdings nicht repräsentativen, ICD-10-Testdatensatz zu erstellen.

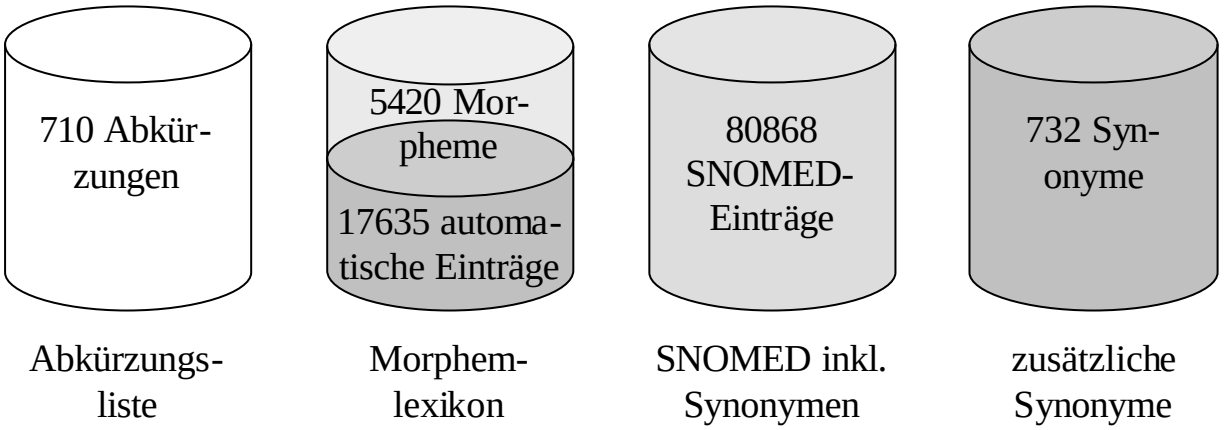
Auf eine Aufnahme der ICD-Einträge mit nicht maximaler Notation, insbesondere von vielen Dreistellern, für die eine weitere Unterteilung existiert, wurde beim Erstellen des Lerndatensatzes ebenfalls verzichtet. Daraus ergibt sich, dass ein Retrieval nach übergeordneten Begriffen über den Umweg der maximalen (im Allgemeinen vierstelligen) Notation erfolgen muss. Wie für die sechsstellig codierten OPS-301-Einträge liegen für diese Klassen keine Lerndaten außerhalb des systematischen Verzeichnisses vor. Vor allem die Alphabethik indes stellt

quantitativ den Großteil der für das Retrieval verwendeten Datensätze, die unter praktischen Gesichtspunkten außerdem der Diagnosenstellung des Arztes näher stehen, als die in einer auf eine Klassifikation zugeschnittenen Weise formulierten Einträge der Systematik mit Zusätzen wie „sonstig“, „nicht näher bezeichnet“ oder „bei Zuständen, die anderweitig klassifiziert sind“ und vielem mehr.

Wie in 4.1 erläutert, müssen alle Einträge, die den Suchraum für eine ICD- bzw. OPS-301-Klassifikation repräsentieren, auch in SNOMED-indexierter Form vorliegen. Dabei werden bis auf die Reihenfolge identische Indexierungen von Lerndaten einer einzigen ICD-Klasse, die sehr häufig dadurch entstehen, dass synonyme Bezeichnungen durch SNOMED auf gleiche Schlüssel abgebildet werden, zu einem einzigen Eintrag zusammengefasst. Liegen identische Indexierungen der Lerndaten unterschiedlicher ICD-Klassen vor, ist die Bedingung der Injektivität verletzt, da die Zuordnung nicht umkehrbar eindeutig ist; dieser Fall tritt auf, wenn Lerndaten nicht exakt genug formuliert sind oder versehentlich manuell fehlklassifiziert wurden, häufigere Ursache ist jedoch eine inkorrekte SNOMED-Indexierung durch MedSearch selbst, wenn ein spezieller Text dem Verschlüsselungsalgorithmus Probleme bereitet oder die zu Grunde liegenden Daten zur Vorverarbeitung, morphologischen Analyse oder SNOMED-Verschlüsselung nicht korrekt sind. In diesem Fall werden die Lerndaten, die zur gleichen SNOMED-Repräsentation führen, bis auf eine ICD- bzw. OPS-301-Klasse gestrichen, der Systemadministrator wird durch eine Rückmeldung auf das Problem aufmerksam gemacht.

Zusätzlich zu den beschriebenen Verzeichnissen, die den Suchraum für das MedSearch-Retrieval aufspannen, stehen dem Algorithmus im Wesentlichen vier weitere Lexika zur Verfügung (Skizze 5.3.2), deren Pflege erforderlich ist. Für die Vorverarbeitung wird eine Liste von 710 *Abkürzungen* verwendet. Die morphologische Analyse basiert auf einer selbsterstellten Liste von aktuell 5420 *Morphemen*, zu der weitere 17635 feststehende Begriffe bei der Zerlegung von SNOMED und obigen Lerndatensätzen automatisch hinzugefügt wurden, da sie durch die 5420 Morpheme nicht zerlegbar sind (großteils Eigennamen und Begriffe aus dem ferneren Umfeld der Medizin). Schließlich arbeitet der eigentliche Verschlüsselungsalgorithmus mit der Nomenklatur *SNOMED*, die 30348 Codes und 80868 Einträge umfasst, also zu einem großen Teil auch Synonyme und verwandte Begriffe, und zu der sich eine Lexemachse gesellt, die der Liste von Lexemen des Morphemlexikons entspricht. Das vierte Lexikon mit weiteren *Synonymen* umfasst 732 Einträge, die zum großen Teil morphologische Verwandtheiten abdecken, die von der morphologischen Analyse selbst nicht erfasst werden

(unterschiedliche Schreibweise des Morphems in Verb und Substantiv; in das Morphemlexikon wegen zu geringer Spezifität nicht aufgenommene Einträge).



Skizze 5.3.2: Weitere MedSearch zu Grunde liegende Lexika

Alle diese Lexika liegen auch in präcodierter Form vor: Das Abkürzungsverzeichnis ist anhand der vor der Auflösung von Abkürzungen stehenden Vorverarbeitungsschritte normiert; die Einträge des Morphemlexikons stehen in vorverarbeiteter Form zur Verfügung; sämtliche SNOMED-Terme wurden morphologisch analysiert; alle Begriffe des Synonymlexikons (jeder Eintrag umfasst zwei äquivalente Begriffe) schließlich müssen nach SNOMED indexiert sein. Auch bei der Präcodierung dieser vier Datensätze werden doppelt vorkommende Projektionen registriert, so dass entweder das fehlerhafte Original entfernt werden kann oder der Algorithmus bzw. die zur Codierung verwendeten Lexika geändert werden.

Die Struktur der Abkürzungsliste wurde bereits in 3.2 definiert; die Nomenklatur SNOMED ist, mit Ausnahme der sich aus dem Morphemlexikon ergebenden L-Achse, vorgegeben, und der zusätzliche Synonymdatensatz besteht aus einer einfachen Ersetzungstabelle. Einen Sonderstatus hingegen nimmt das Morphemlexikon ein. Da die darin vorkommenden Einträge den Algorithmus in komplexerer Form beeinflussen, eine Verwaltung und Erweiterung des selbst erstellten Lexikons durch den Systembetreuer aber dennoch möglich und durchschaubar sein muss, befassen sich die nächsten beiden Abschnitte mit Hinweisen zu seiner Pflege, seiner Datenstruktur und seinem Wachstumsverhalten.

5.4 Verwaltung und Speicherung des Morphemlexikons

Wie bereits in [Schulz 00] beschrieben, wurde für die interaktive Verwaltung des Morphemlexikons eine eigenständige Benutzeroberfläche implementiert, die in Abbildung 5.4.1 zu sehen ist. Mit ihr steht insbesondere eine Plattform für das Erstellen des Lexikons zur Verfügung, für das zur Beginn des Projekts noch kein Datenmaterial vorhanden war. Dabei wurde zunächst der Algorithmus der morphologischen Analyse auf Basis einer leeren Morphemtabelle implementiert, die dann in Abhängigkeit von den Resultaten der Zerlegung (zu Beginn ist keine Zerlegung möglich) einer Liste von Begriffen des medizinischen Wortschatzes schrittweise durch Morpheme aufgefüllt werden konnte. Das Anwachsen des Morphemlexikons wurde dabei in einer kleinen Studie untersucht (Abschnitt 5.5).

Die Begriffsliste, die für das Erstellen der Morphemliste herangezogen wurden, subsummiert dabei sinnvollerweise sämtliche Wörter aus dem Lerndatensatz für das Retrieval (ICD-/OPS-301-Systematik, Alphabet und Ergänzungen), der Synonymtabelle und der Nomenklatur SNOMED. Mit der Plattform kann der jeweils nächste Begriff angefordert werden, wahlweise auch nur, wenn er noch nicht morphologisch zu analysieren ist, oder in einer bestimmten Reihenfolge wie z.B. nach der Länge, da umfangreichere Begriffe im Allgemeinen aus vielen Morpheme zusammengesetzt sind.

In jedem Trainingszyklus wird nun ein neuer Begriff aus der Tabelle zerlegt oder durch das Hinzufügen von Morphemen zerlegbar gemacht. Ein Morphem kann entsprechend dem Wortmodell von MedSearch (Abschnitt 3.3) definiert werden als Wortstamm, Präfix (Vorsilbe), Infix (Fugenmorphem), Derivationssuffix (Nachsilbe), Flexionssuffix (Deklination) oder als fester Begriff (Eponym und Funktionswort bzw. Stopwort). Ein semantisches Gewicht zwischen 0 und 2 (Definition in Abschnitt 3.3) wird dabei jeweils vorgeschlagen (Wortstamm/Eponym/Stopwort 2, Vorsilbe/Nachsilbe/Flexion 1, Fugenmorphem 0), kann aber verändert werden, wenn beispielsweise eine Nachsilbe wie -itis („Entzündung“) von der Bedeutung her einem Wortstamm entspricht. Die Art und Gewichtung eines Morphems kommt in der Schreibweise der Zerlegungen zum Ausdruck; Morpheme mit dem Gewicht 2 werden in Großbuchstaben dargestellt, alle anderen in Kleinbuchstaben. Die Zuordnungsmöglichkeit der Morpheme zu einer Sprache ist für künftige Weiterentwicklungen des Algorithmus vorgesehen; allerdings können Teile eines zusammenhängenden medizinischen Terminus durchaus von verschiedenen Sprachen abstammen (z.B. im Begriff „Diabetes mellitus“, in dem die Namensgebung auf das griechische Wort „diabainein“ zurückgeht, das „Hindurchgehen“ be-

deutet und als ein Synonym für den verstärkten Harnfluss steht, wohingegen der Begriff „mellitus“ lateinischer Abstammung ist und sich auf den „honigsüßen“ Geschmack des Harns bezieht). Morpheme, die fehlerhaft in das Lexikon aufgenommen wurden oder erst im Folgenden erkennbare Probleme verursacht haben, können gelöscht werden. Wenn die Veränderungen am Morphemlexikon abgeschlossen sind, muss eine Neuinitialisierung der Datenbank stattfinden, da sämtliche mit Hilfe des Morphemverzeichnisses zerlegten Tabellen – SNOMED, die Synonymliste, ICD-9/-10 und OPS-301 – nun neu morphologisch analysiert werden müssen, da die zu klassifizierenden Dokumente sonst in veränderter Form zerlegt werden und mit den Einträgen dieser Datensätze nicht mehr in Einklang stehen. Die Neuverschlüsselung aller Datensätze nimmt eine beachtliche Rechenzeit in Anspruch, so dass Änderungen am Morphemlexikon nur in wichtigen Fällen und en bloc durchgeführt werden sollten.

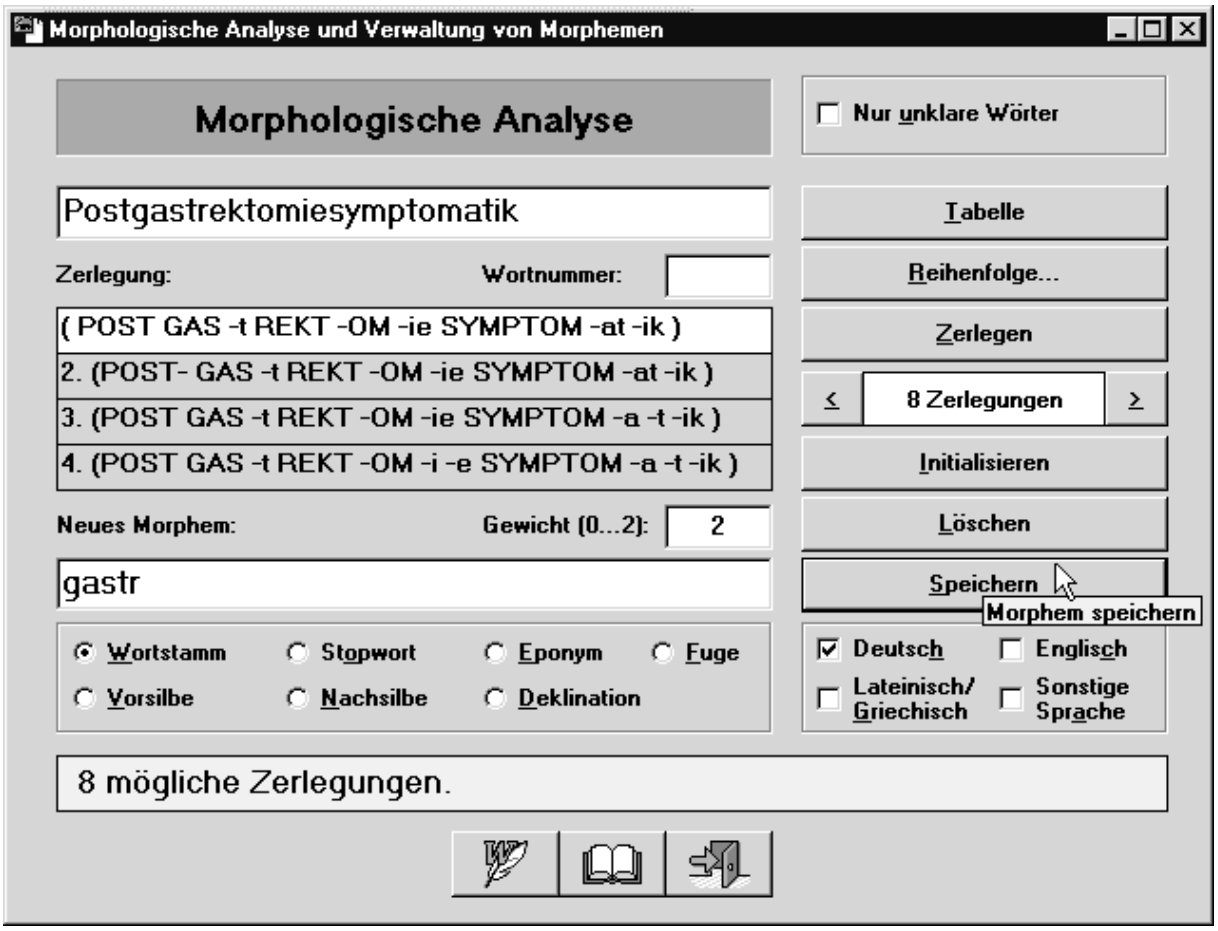


Abbildung 5.4.1: Die Plattform für das Bearbeiten des Morphemlexikons

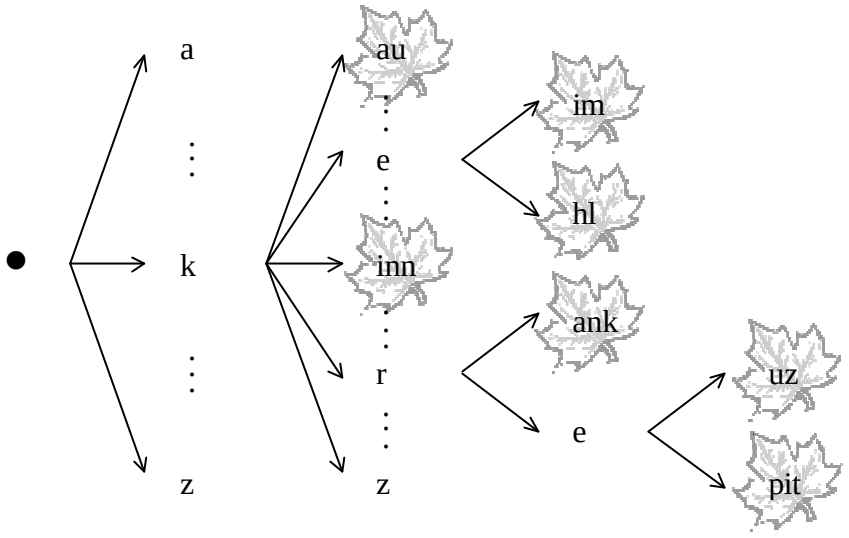
Abbildung 5.4.1 illustriert die Arbeitsweise der Plattform für das Morphemlexikons am Beispielbegriff „Postgastrektomiesymptomatik“, der nicht korrekt analysiert wird, da das Mor-

phem „gastr“ dem Lexikon erst noch hinzugefügt werden muss (es wurde für das Beispiel entfernt). Dennoch sind in diesem Fall acht andere Zerlegungen möglich, die in der anhand der Disambiguierungskriterien aus 3.3 ermittelten Reihenfolge ausgegeben werden. Durch die Hinzunahme von neuen Morphemen, in diesem Fall „gastr“, wächst die Fähigkeit des Algorithmus zunehmend, Begriffe zu segmentieren.

Eine kurze Erwähnung wert ist die praktische Umsetzung der Speicherung von Morphemen, die dem späteren raschen Wiederfinden dienen soll. Dazu wurden in der ersten Fassung des MedSearch-Algorithmus die Morpheme, die zur Wortzerlegung notwendig sind, in einer Pointerstruktur gespeichert. Eine solche Zeigerstruktur hat den prinzipiellen Vorteil, dass eine Suche nach einem Morphem sehr viel gezielter und rascher erfolgen kann, da pro Anfrage bei einem Wort der Länge l maximal l Zugriffe notwendig sind, während die Suche in einer linearen, alphabetisch angeordneten Liste mit n Morphemen (mit den festen Begriffen mehrere 10.000) in der Größenordnung von ld(n) Zeichenkettenvergleichen liegt. –. Da das aktuell verwendete Datenbanksystem ACCESS solche Pointer allerdings nicht unterstützt, wurde inzwischen von dieser Vorgehensweise abgegangen; die Suche nach Morphemen erfolgt nun über einen Index. Trotzdem wird das Verfahren bei Datenbanken und Programmierumgebungen mit Pointerunterstützung eine effizientere Suche gestatten, was vor allem beim Zerlegen von großen Mengen an Dokumenten eine relevante Zeitersparnis bedeuten kann, das für eine mitunter erforderliche morphologische Analyse des umfangreichen Lerndatensatzes und der SNOMED-Einträge oder zur automatischen Klassifikation großer Mengen von Diagnosen zu Test- oder statistischen Zwecken notwendig ist. Daher sei dieses Konzept hier vorgestellt.

Um eine große Menge an Morphemen rasch durchsuchen zu können, ist es sinnvoll, sie in einer Art Baumstruktur zu speichern. Diese Struktur besteht aus je einer Wurzel entsprechend den sieben Morphemarten Wortstamm, Präfix, Suffix, Infix, Flexion, fester Begriff und Zahl. Jeder Buchstabe eines Morphems bezeichnet nun einen Knoten, der auf 26 weitere Buchstaben bzw. auf ein Symbol für Morphemende zeigen kann. Weil es allerdings nicht sinnvoll ist, für jeden Buchstaben aller Morpheme einen eigenen Knoten im Baum (mit 26 weiteren Zeigern) zu reservieren, verweist die letzte Verzweigung eines Astes immer auf eine Liste mit Morphemenden (Blättern). – In ihren Eigenschaften entspricht eine solche Baumstruktur im Wesentlichen den sogenannter *B-Trees*; allerdings ist hier nicht gewährleistet, dass bis auf die Wurzel jeder Knoten auf mindestens $\lceil n/2 \rceil$ (n = Anzahl der Zeiger pro Knoten = 26) weitere Knoten verweist, und die Blätter des Baumes kommen auf sehr unterschiedlichen Leveln zu liegen.

Skizze 5.4.2 illustriert diese Erläuterungen anhand der Baumstruktur eines Morphemlexikons, das die Wortstammliste {kau, kehl, keim, kinn, krank, kreuz, krepit} enthält; dabei sind die Morphemenden durch Blätter des Baumes symbolisiert.



Skizze 5.4.2: Speicherung von Morphemen mittels Baumstruktur

Für die Umsetzung der Analyse von Wörtern anhand des Wordmodells aus Abschnitt 3.3, die alle möglichen Zerlegungen eines Wortes ermitteln soll, ist dabei notwendig, eine Suchanfrage *s* an die Baumstruktur zu stellen, die alle Morpheme ermittelt, die den Anfang dieses Wortes abdecken. So ergäbe die Anfrage *s*(„darmbeinkamm“) an die Wurzel des Baumes, der die Wortstämme enthält, die möglichen Morpheme „darm“ und „darmbein“, da beide im Lexikon enthalten sind. Der Algorithmus würde in der Folge rekursiv die weitere Zerlegung von „beinkamm“ bzw. „kamm“ ermitteln wie in Abschnitt 3.3 beschrieben.

Für den Zeichenkettenvergleich, ob in Baumstruktur oder anderer Form, sind dabei die erläuterten Ausnahmeregeln zu beachten, die für eine hochwertige morphologische Analyse hilfreich sind, die Verarbeitung indes erschweren. So wird, falls keine Zerlegung möglich ist, beim Anfragebegriff ae auf a, oe auf o bzw. ue auf u abgebildet. Damit passt nun „laus“ in „laeuse“, „person“ in „persoenlichkeit“ usw. Des Weiteren kann ein c am Ende eines gespeicherten Morphems sowohl für k als auch für z stehen. „ulc“ beispielsweise findet sich sowohl in „ulkus“ als auch in „ulzera“ usw. Dies wird dadurch bewerkstelligt, dass im Baum sowohl „ulk“ als auch „ulz“ mit dem gleichen Morphemidentifier gespeichert werden. Des weiteren

darf an Stellen, die im Anfragebegriff mit eckigen Klammern gekennzeichnet sind („[ae]“, „[ss]“) nicht getrennt werden, da diese Zeichenfolgen ursprünglich für ein einzelnes Zeichen standen und eine morphologische Trennung innerhalb eines Zeichens nicht denkbar ist. – All die beschriebenen Methoden müssen auch für den Vergleich der Morphemenden, also der „Blätter“ des Suchbaumes gelten.

Die Umsetzung der Baumstruktur wird durch all die Sonderregelungen beim Vergleich von Zeichenketten etwas komplex, doch bleiben diese Probleme bei anderen Arten der Morphemspeicherung wie etwa mittels indexierte Tabelle, die keine so rasche Suche ermöglichen, bestehen. Zusätzlich müssen Befehle für das Fortschreiben des Baumes, das sich einfach darstellt, und für das Löschen von Elementen, das einem verzweigteren Algorithmus folgt, vorgesehen werden. Das Einlesen einer Baumstruktur aus einer linearen Tabelle würde bei jedem Start des Systems zudem ein *Preload* notwendig machen. Sie bietet daher die größten Vorteile bei der Wahl von Programmierumgebungen und Datenbanken, die Pointerstrukturen gezielt unterstützen.

5.5 Studie über das Wachstumsverhalten eines Morphemlexikons

Um über den Aufwand für das Erstellen und den zu erwartenden Umfang eines Morphemlexikons, für das im Gegensatz zu den anderen für den präsentierten Ansatz benötigten Lexika keine Daten zur Verfügung standen und das für Fortentwicklungen erweitert oder neu erstellt werden sollte (und zum Teil schon wurde), eine fundierte Abschätzung zu erhalten, werden im Folgenden die Ergebnisse einer eigenen kleinen Studie vorgestellt, die zum Teil bereits in [Schulz 00] eingegangen sind.

Dabei wurden zunächst aus einer Liste von 104.902 am Heidelberger Universitätsklinikum in mehreren Jahrgängen gestellten Kurzdiagnosen, die alle dort vertretenen Fachbereiche abdecken, zufällig 33.000 Diagnosen ausgewählt und sukzessive die zur Zerlegung notwendigen Morpheme einem zunächst leeren Lexikon hinzugefügt. Lediglich eine Reihe von Abkürzungen wurde vorgegeben. Einen kurzen Einblick in den Datensatz aus Heidelberg, der unabhängig von der hier vorgestellten Untersuchung zur Auswertung des MedSearch-Algorithmus herangezogen wurde (Abschnitt 6.2), findet sich in Tabelle 6.2.2. Weder die hier ermittelten Morpheme noch sonstige resultierenden Erkenntnisse über Verbesserungsmöglichkeiten des Algorithmus, dessen Entwicklung zu diesem Zeitpunkt bereits abgeschlossen war, flossen

allerdings in das Experiment in Abschnitt 6.2 mit ein; das Morphemlexikon von MedSearch wurde unabhängig hiervon allein anhand der in SNOMED, in der Synonymliste sowie in den Tabellen des ICD und des OPS-301 vorkommenden Ausdrücke erstellt.

Beibehalten wurden bei der Analyse der Datensätze lyrische Umschreibungen („Angstgetönte depressive Reaktion“ u.dgl.); nicht ins Morphemlexikon eingehen konnten hingegen sämtliche Schreibfehler, ob seriöser („Seriöse Meningitis“) oder eher fröhlicher Natur („Scherzmittelabusus), willkürliche Abkürzungen („Blutng.“ für „Blutung“ u.dgl.) oder Befunde mit diagnostisch recht seltsamen Begriffen („Käsesahnetorte“ u.dgl.); im Zweifelsfalle entschied der Pschyrembel. Insgesamt mussten von den erwähnten 33.000 Diagnosen 2.561 (7,76 %) verworfen werden. Beim Erstellen des Lexikons anhand dieser Daten wurde jeweils die Anzahl an Morphemen ermittelt, die notwendig waren, um die ersten n Diagnosen zu zerlegen (n = 10, 50, 100, ...). Dabei ergab sich folgendes Wachstumsverhalten (ausgewählte Werte):

Diagnosen	Morpheme
10	65
50	214
100	330
250	616
500	919
750	1134
1000	1299
1500	1583
2000	1794
2500	1953
3000	2102
4000	2332
5000	2506

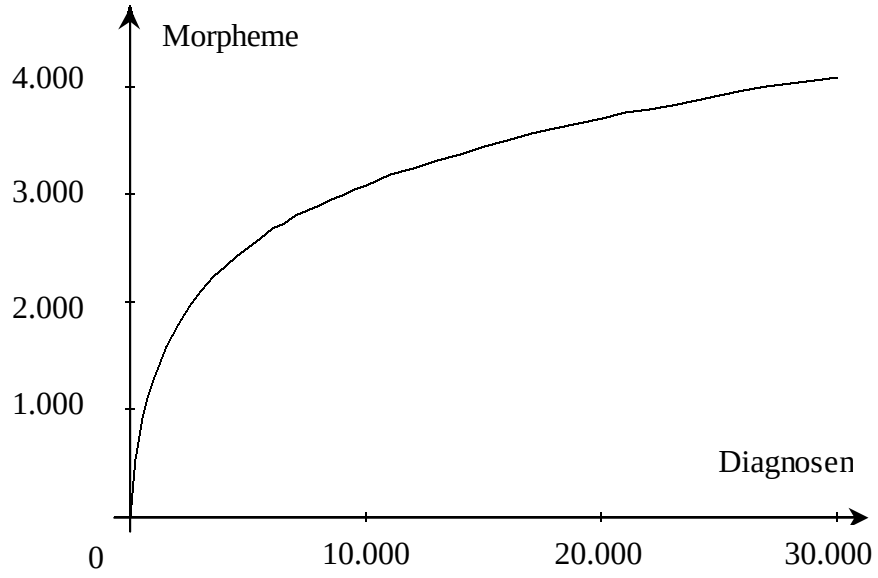
Diagnosen	Morpheme
6000	2679
7000	2803
8000	2900
9000	2992
10000	3085
11000	3181
12000	3243
13000	3311
14000	3369
15000	3449
16000	3499
17000	3568
18000	3614

Diagnosen	Morpheme
19000	3658
20000	3706
21000	3756
22000	3790
23000	3825
24000	3876
25000	3916
26000	3963
27000	3998
28000	4032
29000	4061
30000	4088
30430	4098

Tabelle 5.5.1: Anzahl der Morpheme, die zur Zerlegung klinischer Diagnosen notwendig war

Die graphische Verteilung dieser Werte in Abbildung 5.5.2 zeigt deutlich ein zu erwartendes logarithmisches Wachstumsverhalten der Ordnung $O(\log n)$ sowie eine äußerst geringe Schwankungsbreite der Einzelwerte, so dass für die Größe von Morphemlexika bei Fähigkeit zur Analyse einer umfangreicheren Anzahl an Begriffen zuverlässige Vorhersagen getroffen werden können.

Wichtig ist daneben vor allem die erneute Schlussfolgerung, dass mit einem sehr begrenzten Morphemschatz weiteste Teile des medizinischen Sprachgebrauchs abgedeckt werden können. Allerdings bestätigt die Kurve ebenfalls, dass es nie möglich sein wird, ein solches Lexikon zu komplettieren; von allen neuen Diagnosen, mit denen das System konfrontiert wird, ist a priori ein geringer, aber nicht vernachlässigbarer Prozentsatz nicht zu zerlegen. Der Umfang des Lexikons hängt wesentlich mit davon ab, wie „fein“ in Morpheme zerlegt wird; je atomarer diese Morpheme sind, desto weniger werden notwendig sein. Gleichzeitig steigt damit allerdings die Anzahl an inkorrekten Analysen, die in dieser Untersuchung nicht berücksichtigt wurden. Prinzipiell wird sich jedoch auch bei weniger großzügiger Zerlegung die gleiche Kurve ergeben, die dem Anwachsen jedes Morphemlexikons von vornherein enge Grenzen setzt.



Grafik 5.5.2: Wachstumsverhalten eines Morphemlexikons

6 Evaluation: Zwei Verfahren zur Analyse des Algorithmus und Ergebnisse

Im folgenden Abschnitt 6.1 wird eine erste qualitative Fehleranalyse durchgeführt, die auf Verbesserungsmöglichkeiten in Bezug auf den MedSearch-Algorithmus und die zu Grunde liegende Wissensbasis abzielt. Eine größere, quantitative Studie präsentiert in der Folge Abschnitt 6.2, mit der die aktuelle Leistungsfähigkeit des Verfahrens im Vergleich mit einem konventionellen Ansatz eingestuft und beurteilt wird. Die Diskussion und Interpretation der Resultate erfolgt abschließend in Kapitel 7.

6.1 Orientierende Analyse verbleibender Fehlerquellen

Um den MedSearch-Algorithmus möglichst praxisnah zu entwerfen, wurde bereits im Stadium der Entwicklungsphase eine Liste von 200 Diagnosen aus einem Datensatz der BFA zu Testzwecken herangezogen. Dabei wurde das Feedback genutzt, um eine stufenweise Verfeinerung des Ansatzes durch grundsätzliche, nicht diagnosenspezifische Verbesserungen zu erzielen; viele der beschriebenen Details des Algorithmus haben hier ihren Ursprung. Von den 200 Dokumenten wurden am Ende 68,5 % (Dreisteller) bzw. 55,1 % (Viersteller) übereinstimmend mit der manuellen Codierung durch den Arzt (*gold standard*) nach ICD-9 klassifiziert. (Für die Viersteller konnten dabei nur 138 Einträge ausgewertet werden, da 62 manuell nicht maximal vorcodiert waren.) Für die Fehlklassifikationen dieses quantitativ nicht repräsentativen Datensatzes wurden im Detail die Ursachen ermittelt, auf die im Weiteren eingegangen wird.

Des Weiteren standen drei größere Datensätze zu unabhängigen, quantitativen Testzwecken zur Verfügung. Von einem großen Datensatz an oft überlangen Entlassungsdiagnosen des Uniklinikums Göttingen (1986-1994) wurden willkürlich 1000 ausgewählt und getestet. Aus einer ähnlichen Tabelle an manuell nach ICD-9 codierten Diagnosen vom Uniklinikum Heidelberg, die u.a. viele Spezialabkürzungen enthalten, wurden 10000 Dokumente untersucht (vgl. Abschnitt 6.2). Zusätzlich lag eine Feinaufteilung der ICD-9 in einer klassifikationsspezifischen Sprache (Datensatz aus Dresden) vor, von der wiederum 1000 Einträge zu Testzwecken ausgewählt wurden. Alle drei Datensätze weisen auf Ebene der Dreisteller eine etwa 50 %ige, auf Ebene der vierstelligen Notation etwa 40 %ige Übereinstimmung mit der manuellen Codierung auf. Die Ergebnisse der einzelnen Auswertungen sind in Tabelle 6.1 verzeichnet.

Quelle	BFA	Göttingen	Heidelberg	Dresden
Dreisteller ICD-9	137 (68,5%)	513 (51,3%)	50,4 %	562 (56,2%)
Viersteller ICD-9	76 (55,1%)	400 (40,0%)	39,6 %	381 (38,1%)
Stichprobengröße	200 bzw. 138	1000	10000	1000

Tabelle 6.1.1: Testergebnisse

Die verbleibenden Fehlklassifikationen bei der Codierung des BFA-Datensatzes wurden eingehend untersucht. Die Quellen für die insgesamt 77 nicht mit dem *gold standard* übereinstimmenden Klassifikationen (auch nur in der vierten Stelle) konnten dabei anhand des Protokolls (Abschnitt 5.2) jeweils den Testdaten, den Lerndaten oder dem Algorithmus zugeordnet werden, mit folgenden Anteilen (Prozentangaben bezogen auf sämtliche 200 Diagnosen):

Fehlerquelle	**Anzahl**	**Ursache**
Testdatensatz	23	Falsche ICD-Klasse (*gold standard*) manuell vorgegeben
(insgesamt 28	4	Diagnose unvollständig/uninterpretierbar
Fälle = 14 %)	1	Dokument enthält Diagnosen aus verschiedenen ICD-Klassen
Lerndatensatz	14	Keine vergleichbarere Diagnose im Lerndatensatz
(insgesamt 39	7	Verallgemeinerung notwendig, durch SNOMED nicht abgedeckt
Fälle = 19,5 %)	5	Hierarchie von Krankheitsbegriffen fehlt
	4	Korrektur des Morphemlexikons erforderlich
	3	Fehlender Eintrag im Synonymlexikon
	2	Dokument im Lerndatensatz mehrdeutig/fehlerhaft
	2	SNOMED-Synonyme identifizieren verschiedene ICD-Klassen *(„Struma“ und „Struma simplex“, 4. Stelle)*
	1	Fehlerhafte Auflösung einer Abkürzung
	1	Abkürzung weder auflösbar noch im Lerndatensatz vorhanden
Algorithmus	4	Fehlendes „sonstige/nicht näher bezeichnete“-Konzept (4. Stelle)
(insgesamt 10	2	Krankheitsbegriff fälschlicherweise an 1. Stelle gesucht
Fälle = 5 %)	1	Fehlerhafte Auflösung einer Abkürzung durch den Algorithmus
	1	Fehlcodierung durch überflüssige Information in der Diagnose
	1	Auflösen von „c“ *(„Uricämie“ nicht als „Urikämie“ erkannt)*
	1	Weitere Verallgemeinerung nach SNOMED fand nicht statt

Tabelle 6.1.2: Typische Fehlerquellen bei der MedSearch-Klassifikation

Eine solche Fehleranalyse wäre erst auf Basis einer größeren Datenmenge und einem von der Entwicklung des Systems unabhängigen Testdatensatz repräsentativ, sie muss indes manuell erfolgen und erfordert einen gewissen Zeitaufwand. Eine Tendenz zeichnet sich jedoch bereits ab und wird beim orientierenden Durchsehen von Fehlklassifikationen der Göttinger, Heidelberger und Dresdener Datensätze bestätigt. Die Erläuterung und Interpretation der hier festgehaltenen Ergebnisse erfolgt in Abschnitt 7.1.

6.2 Experimenteller Vergleich von MedSearch mit einem klassischen Ansatz

Von zentraler Bedeutung für die Einschätzung eines neuen Verfahrens ist im Sinne der EBM (*evidence based medicine*) der kritische Vergleich mit anderen, herkömmlichen Alternativen. Um die Leistungsfähigkeit des MedSearch-Algorithmus zu dokumentieren, wurde daher dieser Ansatz anhand eines umfangreichen Datensatzes getestet und mit einem konventionellen Verfahren verglichen. Wesentlich für die Beurteilung der Ergebnisse ist zusätzlich die getrennte Untersuchung der beiden Teile Indexierung und Retrieval. Im Folgenden werden der Testdatensatz, der Aufbau des Experiments sowie die Ergebnisse der Evaluation vorgestellt. Die Resultate dieser Untersuchung und die zentralen Folgerungen daraus wurden bereits in [Franz 00] veröffentlicht.

Der Testdatensatz

Um den Ansatz von MedSearch mit einem Standardverfahren zu vergleichen, wurde aus einem Satz von über 100.000 Entlassungsdiagnosen, die über mehrere Jahre am Uniklinikum Heidelberg von Ärzten aller vertretenen Fachabteilungen codiert wurden, eine zufällige Stichprobe der Größe 10.000 ausgewählt. Für den klassischen Ansatz wurde diese Probe noch einmal auf 2.251 Diagnosen reduziert (Skizze 6.1.1). Diese Einschränkung wurde gemacht, da die konventionelle Methode eine Distanzberechnung zu jedem einzelnen Element des Suchraumes durchführt und daher im Vergleich zum MedSearch-Verfahren drastisch mehr Rechenzeit beansprucht. – Die Daten selbst wurden manuell von den behandelnden Ärzten nach ICD-9 verschlüsselt (*gold standard*); bei der Durchführung des Experiments lagen vergleichbar umfangreiche Datensätze mit ICD-10-Codierung noch nicht vor, wie sie inzwischen existieren. Die Entlassungsdiagnosen wurden in ihrer ursprünglichen Form als nicht vorverarbei-

Skizze 6.2.1: Der Testdatensatz

tete Freitexteingabe gespeichert. Bei der Auswahl der Stichprobe wurden lediglich solche Diagnosen ausgeschlossen, die wegen Überlänge beim Speichern gekappt worden waren (über 56 Zeichen).

Größtes Problem bei der Beurteilung des Datensatzes ist die Kumulierung von Diagnosentexten mit komplett identischer Zeichenfolge, die die wahre Häufigkeit des Auftretens bestimmter Einträge verschleiert. In einem vergleichbaren Satz von Entlassungsdiagnosen des Uniklinikums Göttingen wurde diese Anzahl berücksichtigt; bei ihrer Analyse wird deutlich, wie eine Akkumulation das Spektrum hin zu umständlichen und fehlerhaften Formulierungen verschiebt. Ein typisches Beispiel wären die vier Einträge *Dialatative Kardiomyopathie* (2 mal), *Dialtative Kardiomyopathie* (2 mal), *Dilatative Kardiomopyathie* (1 mal) und *Dilatative Kardiomyopathie* (613 mal), die im Heidelberger Testdatensatz als gleichwertig betrachtet würden. Bei der Beurteilung der Ergebnisse des Experiments wird daher das Verhältnis der einzelnen Verfahren zueinander entscheidender sein als die Größenordnung der korrekten Klassifikationen an sich.

Rechtschreibfehler waren in diesem Testdatensatz sehr häufig, sicher neben der genannten kumulativen Verteilung sowohl durch Zeitdruck als auch mangelndes Interesse der Ärzte an dokumentatorischen Aufgaben zu begründen. Häufig beschränken sich die Einträge nicht auf zu verschlüsselnde Hauptdiagnosen, sondern enthalten viele zusätzliche, für die Klassifikation irrelevante und zum Teil irreführende Informationen. Erschwerend wirken des Weiteren eine große Anzahl willkürlicher Abkürzungen und krankenhausspezifischer Fachausdrücke. Beim Begutachten einer Stichprobe der Daten schienen rund 10 % der von den Ärzten angegebenen Codes, die in unserem Experiment als *gold standard* festgelegt wurden, falsch oder zumindest fragwürdig. Die wegen ihrer Praxisnähe sehr realistischen Testdaten wurden indes verwendet, weil keine Alternative in bedeutend besserer Qualität und vergleichbarem Umfang existierte. – Einen repräsentativen, nicht selektierten Ausschnitt aus dem Heidelberger Datensatz in der Reihenfolge des Auftretens der willkürlich permutierten Diagnosen bietet die Tabelle 6.2.2.

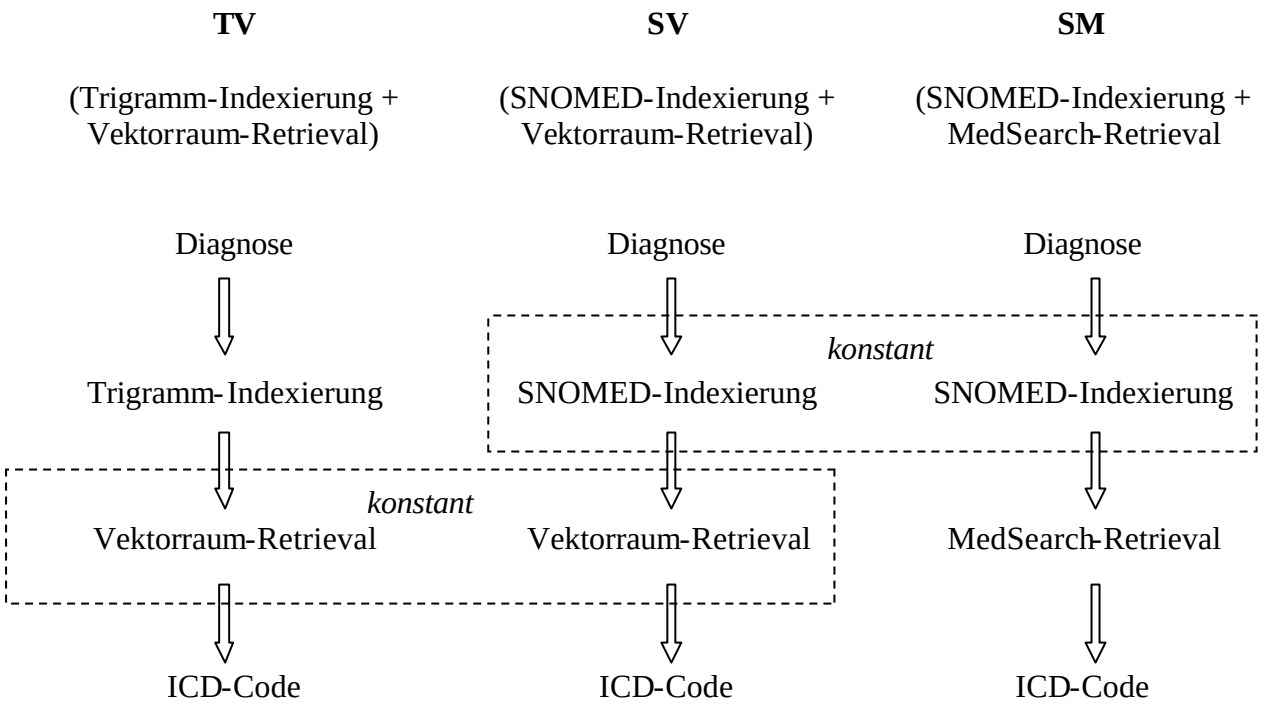
Identifier	ICD-9	Diagnose
104785	344.8	Spast-Tetraparese mit Katheterfehllage bei Bac.-Pumpe
104786	715.9	Arthropathie, degenerative
104787	866.0	Flankentrauma links mit Nierenkontusion
104788	193	papilläres SD-Ca T4N1M0 G II re.
104789	829.0	Z.n. Aitken I - Verletzung links
104790	560.2	Sigma Subvolvolus
104791	415.1	V.a. Lungenembolie mit Lungenoedem
104792	277.3	Amyloidose linke Orbita
104793	174.9	Rekonstruktionswunsch nach Ma-Ca
104794	478.1	KH-Cyste links
104795	320.2	bakterielle Meningitis durch Streptokokkus suis
104796	196.0	LK-Metastasen li. KW
104797	459.0	Z.n. OP-Zahnsanierung bei Risiko (Blutungsneigung)

Tabelle 6.2.2: Der Testdatensatz (Ausschnitt)

Hervorzuheben ist bei der Betrachtung dieser Diagnosentexte, dass sie sich wesentlich unterscheiden von den Angaben, die einem automatischen Codiersystem typischerweise vorgegeben werden und bei denen der Arzt sich auf für die Klassifikation wesentliche Fakten beschränkt, sowie das Verständnis spezieller Abkürzungen nicht voraussetzt.

Aufbau des Experiments

Bei der Durchführung des vergleichenden Experiments wurden folgende drei Ansätze gewählt: Die klassische Trigramm-Indexierung, mit anschließendem Vektorraum-Retrieval; die SNOMED-Indexierung mit anschließendem MedSearch-Retrieval, wie beschrieben; und als dritte, hybride Variante wurden die nach unserem Verfahren SNOMED-indexierten Dokumente mit einem Vektorraum-Retrieval verknüpft. Da das MedSearch-Retrieval offensichtlich nicht auf Basis von Trigrammen arbeiten kann, blieb eine vierte Kombinationsmöglichkeit aus. Die Skizze 6.2.3 illustriert den Versuchsaufbau.



Skizze 6.2.3: Versuchsaufbau

Trigramme. Die N-gramm-Verschlüsselung ist ein einfaches, lexikonfreies und sprachunabhängiges Verfahren für das Indexieren von Dokumenten. Jedes Dokument wird dabei durch einen Index repräsentiert, der aus allen möglichen Zeichenketten der Länge N besteht, die darin enthalten sind (diese Dokumente müssen mindestens N Zeichen umfassen). Trigramme im Besonderen sind N-gramme der Länge N=3. So wird zum Beispiel die Suchanfrage „Hypertension“ beim angewandten Verfahren zum Index (hyp, ype, per, ert, rte, ten, ens, nsi, sio, ion). Zu diskutieren wäre hier die Hinzunahme von Trigrammen am Anfang und Ende des Dokumentes, die Leerzeichen enthalten (hier --h und -hy bzw. on- und n--), da diese in der aktuellen Form etwas unterbewertet werden.

Vektorraum-Retrieval. Der zu einem Dokument gehörende Index, der aus einer Menge von Termen (zum Beispiel Trigrammen) $t_1 ... t_n$ besteht, kann als Element eines Vektorraums aufgefasst werden, der für jeden denkbaren Term der (endlichen) Grundmenge eine eigene Dimension besitzt. Dieser Vektor kann verkürzt als $t = (t_1, w_1; t_2, w_2; ... ; t_n, w_n)$ dargestellt werden mit verschiedenen Gewichten $w_1 ... w_n$ für jede der n Dimensionen $t_1 ... t_n$ (die Länge des Vektors in jeder anderen Richtung beträgt 0). In die Gewichte für die einzelnen Vektoren, die größer als 0 sind und maximal 1 betragen, gehen dabei die Häufigkeit eines Terms in einem

Dokument und die inverse Häufigkeit des Dokuments selbst mit ein. Durch Normalisierung der Gewichte wird zudem die Länge der einzelnen Dokumente berücksichtigt, da beim Retrieval längere Dokumente sonst bevorzugt werden.

Auf gleiche Art wie die Elemente des Suchraumes müssen auch die Suchanfragen selbst indexiert und den einzelnen Indizes ihre Gewichte zugeordnet werden. Nun kann durch einen Vergleich des Vektors der Suchanfrage Q (*query*) mit einem Dokument D des indexierten Suchraumes die Ähnlichkeit von Anfrage und Dokument bestimmt werden. Dafür wird das innere Produkt zwischen den normalisierten (d.h. auf Länge 1 normierten) Vektoren berechnet:

$$d(\vec{D},\vec{Q}) = cos(\vec{D},\vec{Q}) = \frac{\vec{D}}{|\vec{D}|} \times \frac{\vec{Q}^T}{|\vec{Q}|}$$

Dieses innere Produkt entspricht bei normalisierten Vektoren dem Cosinus des Winkels zwischen Q und D (Cosinusmaß), d.h. einem Wert zwischen 0, wenn die beiden Vektoren senkrecht zueinander stehen (d.h. wenn keine der Indizes übereinstimmen) und 1, wenn die Dokumente identisch sind.

Für den so definierten Vektorraum wird nun (unter brachialem Rechenaufwand) ein Retrieval definiert, indem das Cosinusmaß zwischen der Anfrage und sämtlichen Dokumenten des Suchraums berechnet wird und diese in eine Reihenfolge gebracht werden, in der diese Vektorraumprodukte absteigend angeordnet sind. So ergibt sich nicht nur das der Anfrage im Sinne der Vektorraummetrik nächstgelegene Element, sondern eine gerankte Liste, aus der der Benutzer bei der semiautomatischen Klassifikation das gewünschte Dokument auswählen kann.

Dieses Vektorraum-Retrieval, das auf G. Salton zurückgeht, ermöglicht eine einfache Berechnung eines symmetrischen Maßes zwischen Suchanfrage und Dokumenten des Suchraumes. Als Gewichte werden dabei in unserem Experiment die relative Häufigkeit der Bezeichner im indexierten Suchraum verwendet, der bei allen zu untersuchenden Varianten identisch war (der in Abschnitt 5.3 beschriebene Lerndatensatz). Dass bei der Definition des Vektorraum-Retrievals und der Gewichtung die zu Grunde liegenden Indexterme unterschiedlicher Art sein können, wurde für den Versuchsaufbau genutzt, indem wie in Skizze 6.2.3 dargelegt zum Vergleich zwischen klassischem Ansatz (Trigramm-Indexierung mit Vektorraum-Retrieval)

und dem hier vorgestellten Verfahren (SNOMED-Indexierung und MedSearch-Retrieval) eine gemischte Variante (SNOMED-Indexierung und Vektorraum-Retrieval) hinzugefügt wurde, die es in der Folge erst ermöglicht, die Ursache für den Qualitätsunterschied zwischen den beiden konkurrierenden Ansätzen genauer zu lokalisieren.

Detailliertere Informationen über das gängige Vektorraum-Retrieval bieten [Salton 91, Salton 94, Wiesman 97].

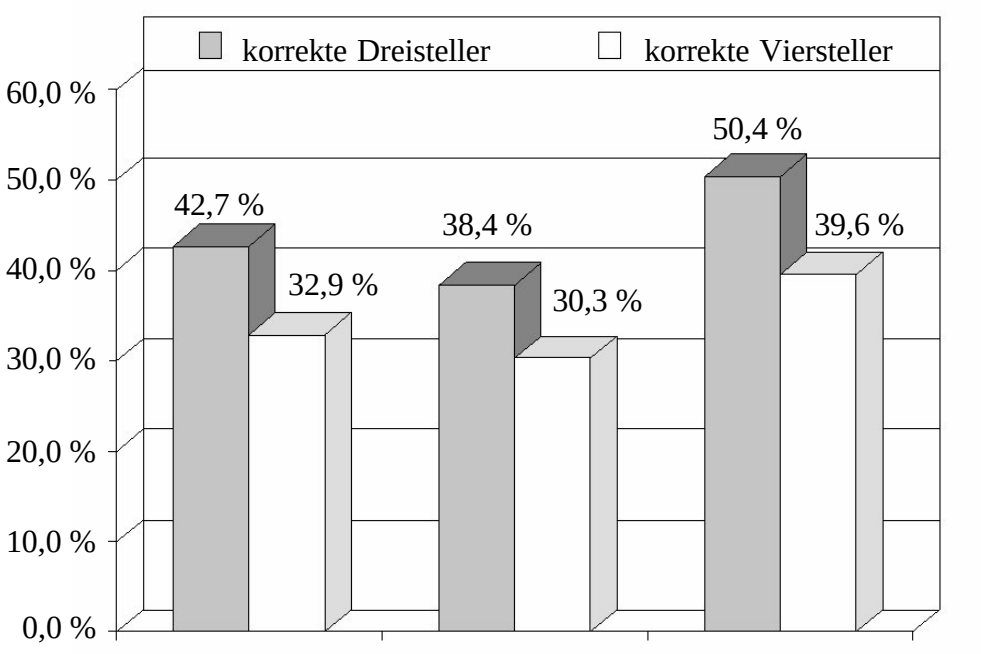
Ergebnisse

Die Ergebnisse der automatischen ICD-9-Codierung des Testdatensatzes mit den beschriebenen drei Methoden sind in Tabelle 6.2.4 festgehalten; neben dem ermittelten Anteil an korrekten (d.h. mit dem *gold standard* übereinstimmenden) Klassifikationen wurde die benötigte Rechenzeit dokumentiert. Eine ausführliche Diskussion der Werte erfolgt in Abschnitt 7.2.

	TV	SV	SM
Indexiermethode	Trigramme	SNOMED	SNOMED
Retrievalmethode	Vektorraum	Vektorraum	MedSearch
n	2.251	10.000	10.000
korrekte Dreisteller [95%-Konfidenzindervall]	42,7 % [40,6 % - 44,8 %]	38,4 % [37,5 % - 39,4 %]	50,4 % [49,4 % - 51,3 %]
korrekte Viersteller [95%-Konfidenzindervall]	32,9 % [31,0 % - 34,9 %]	30,3 % [29,4 % - 31,2 %]	39,6 % [38,7 % - 40,6 %]
Rechenzeit pro Diagnose (Pentium 2 133 MHz)	53,495 s	1,605 s	0,973 s

Tabelle 6.2.4: Resultate der experimentellen Untersuchung

Die graphische Darstellung in Diagramm 6.2.5 soll die experimentell ermittelten Ergebnisse und die Unterschiede zwischen den Resultaten der drei untersuchten Ansätze illustrieren.



Indizes:	Trigramme	SNOMED-Codes	SNOMED-Codes
Retrieval:	Vektorraum	Vektorraum	MedSearch
	(n = 2.251)	*(n = 10.000)*	*(n = 10.000)*

Diagramm 6.2.5: Resultate in graphischer Darstellung

7 Diskussion und Ausblick

Im letzten Teil dieser Arbeit werden die in Kapitel 6 durchgeführten Untersuchungen zur Beurteilung des MedSearch-Algorithmus ausgewertet und dabei die Vor- und Nachteile des Verfahrens sowie Ansätze für künftige Weiterentwicklungen erörtert. Abschnitt 7.1 schließt sich an die Ergebnisse aus 6.1 an, bei denen es in erster Linie um Probleme und Verbesserungsmöglichkeiten der präsentierten Methode geht. In Abschnitt 7.2, der die Resultate des Experiments aus 6.2 näher erläutert, soll der Vergleich mit dem klassischen Ansatz interpretiert und diskutiert werden. Eine resümierendes Fazit in Abschnitt 7.3 fasst die wichtigsten Resultate der Arbeit noch einmal zusammen und schließt sie mit einem Ausblick auf mögliche Weiterentwicklungen ab

7.1 Diskussion der Analyse verbleibender Fehlerquellen

Die Ergebnisse der Untersuchung in 6.1 lassen Rückschlüsse auf die Ursachen von Fehlerquellen bei der automatischen Indexierung und Klassifikation von Diagnosentexten zu, die zum Teil grundsätzlicher Art (Testdaten, Lerndaten), zum Teil durch den MedSearch-Algorithmus selbst bedingt sind.

Von den klinischen **Testdaten** muss ein deutlicher Anteil (in der Praxis um 10%) als inkorrekt codiert betrachtet werden. Da diese durch den Arzt vorgenommene Klassifikation für den Test als *gold standard* verwendet wurde, kann davon ausgegangen werden, dass ein erheblicher Prozentsatz der als falsch eingestuften Ergebnisse der automatischen Klassifikation in der Tat die korrekte ICD-Klasse widerspiegelt. Dadurch erhöht sich der Anteil an korrekten Ergebnissen von den bei den verwendeten Testdatensätzen etwa 50 % auf real etwa 56 %; gleichzeitig beträgt der Prozentsatz an richtigen Klassifikationen bei manueller Bearbeitung nur um 90 %. Die Diskrepanz zwischen der Qualität automatischer und manueller Klassifikation liegt bei Entlassbriefdiagnosen also bei etwa 34 %. Allerdings sind maschinelle Fehlklassifikationen oft bedeutend sinnentstellender, also von der wahren Diagnose semantisch entfernt, als manuelle. Dies liegt daran, dass der Arzt bei der Wahl eines falschen Codes meist eine Klasse mit verwandter Bedeutung wählt, während der Algorithmus seine Auswahl großteils nach morphosyntaktischen Ähnlichkeiten trifft.

Typische Probleme des Testdatensatzes sind häufige Schreibfehler, die Angabe von zahlreicher für die Klassifikation überflüssiger Information, die Verwendung von klinikumsinternen Begriffen und Abkürzungen, oder das willkürliche Kürzen von Wörtern. Selten genügt ein Diagnosentext inhaltlich nicht, um die korrekte Klasse zuzuordnen; häufiger enthält er unterschiedliche Einzeldiagnosen nebeneinander, die nur durch mehrere ICD-Klassen abzudecken sind.

Dadurch, dass mehrfach vorkommende Entlassungsdiagnosen im Testdatensatz nur einfach registriert wurden, wird das Spektrum an Einträgen stark verzerrt von häufigen und korrekt zugeordneten Texten hin zu selteneren und fehlerhaft geschriebenen Diagnosen, die der automatischen Klassifikation Probleme bereiten. Nach dem klassischen *Zipfschen Gesetz* (George Kingsley Zipf, 1902-1950) [Knüppel 01], angewandt auf die Häufigkeit von Diagnosentexten, machen die 10 % häufigsten Befunde bereits ungefähr 90 % aller auftretenden Befundstellungen überhaupt aus; dieses Verhältnis wird durch die Kumulation identischer Diagnosentexte zugunsten der weniger häufigen Diagnosen verschoben, für die im Lerndatensatz indes weniger Datenmaterial existiert und die bei der automatischen Klassifikation folglich in aller Regel schlechter abschneiden.

Große Schwierigkeiten finden sich auch auf Ebene der **Lerndaten**; sie betreffen sämtliche der automatischen Klassifikation zu Grunde liegenden Lexika. Nur unvollständig ist derzeit die Liste auflösbarer *Abkürzungen*, sie sollte durch die Auswertung klinischer Datensätze den Bedürfnissen in der Praxis angepasst werden. Im Gebrauch von Abkürzungen bestehen große Unterscheide zwischen den einzelnen Fachabteilungen, so dass ein nach den entsprechenden Bereichen untergliedertes Abkürzungsverzeichnis von deutlichem Nutzen wäre. Des weiteren bestehen auch lokale Unterschiede im Sprach- und Abkürzungsgebrauch verschiedener Kliniken, so dass das System für klinikumsinterne Ergänzungstabellen offen sein sollte.

Das *Morphemlexikon* ist in der aktuell eingesetzten Form sehr unpräzise. Es basiert auf einem ersten Entwurf, dessen Änderung in der Folge zu zeitintensiv und durch die Notwendigkeit, im Anschluss die Nomenklatur SNOMED, die Synonymliste und den ICD-Suchraum neu zerlegen zu müssen, nicht praktikabel erschien. Eine neuere Version wurde für Folgeprojekte bereits umgesetzt. Prinzipiell sollte die Wartung des Morphemlexikons durch ein Feedback anhand der durch das System falsch klassifizierten und morphologisch analysierten Diagnosen erfolgen, sowie Neuerungen des medizinischen Sprachgebrauchs berücksichtigen. Ein gut erstelltes Morphemlexikon ist in der Regel indes wenig wartungsintensiv.

Auf Ebene des *SNOMED* sind Änderungen nicht vorgesehen. Wünschenswert wäre allerdings der Einsatz einer neueren SNOMED-Version mit mehrschichtigeren, ein breiteres Spektrum umfassenden Termen sowie besseren Möglichkeiten für die Hierarchie- und Kategoriebildung. Derzeit liegt eine solche Version in deutscher Sprache allerdings nicht vor. Das Problem der zu weitgehenden Synonymisierung durch SNOMED, das bei weiterer Verfeinerung der Klassifikation durch zunehmend detailliertere Ordnungssysteme an Bedeutung gewinnen könnte, kann durch das Hinzufügen von zusätzlichen Identifiern zu SNOMED in Ausnahmefällen gelöst werden (für die explizit angeforderte SNOMED-Indexierung ist eine Projektion solcher Identifier auf einen gemeinsamen SNOMED-Code problemlos möglich).

Zusätzlich zu den *Synonymen*, welche die SNOMED II bereits enthält, wäre die weitere Synonymtabelle begrenzt auszubauen; allerdings werden auch zunehmend Fälle registriert, bei denen durch eine Gleichsetzung von Begriffen Diagnosentexte fälschlicherweise auf identische Schlüsselfolgen projiziert werden, obwohl sie unterschiedlichen ICD-Klassen zuzuordnen sind. Insbesondere bei der Indexierung des ICD-Lerndatensatzes bereiten zu weitgehende Synonyme Probleme, da Projektionen auf den SNOMED-indexierten Suchraum umkehrbar eindeutig sein müssen. – Eine weitere Analyse der separaten *L-Achse*, die medizinisch relevante, nicht durch SNOMED abgedeckte Lexeme umfasst, wäre wünschenswert und wurde im Rahmen der Neubearbeitung des Morphemlexikons bereits weitgehend durchgeführt. In der bisherigen Fassung enthält sie noch viele sinnvoll zerlegbare Einträge, die großteils automatisch bei der morphologischen Analyse von SNOMED, der Synonymliste und den Tabellen des ICD-Suchraums gesammelt wurden. – Erforderlich wäre des Weiteren eine Taxonomie, welche die *Begriff-Oberbegriff-Relation* ergänzt, die im Algorithmus mittels der Hierarchie von SNOMED ermittelt wurde (Abschnitt 4.5) und die zu unvollständig und ungenau ist. Krankheitsbegriffen sollte ein semantisches Gewicht zugeordnet werden, um beim häufigen Auftreten mehrerer Terme innerhalb eines Diagnosentextes, die eine Krankheit bezeichnen, den für die Wahl der korrekten ICD-Klasse relevanteren feststellen zu können.

Von zentraler Bedeutung bei der Pflege der Lexika ist abschließend die ständige Anpassung und Erweiterung des *ICD- bzw. OPS-301-Lerndatensatzes* (systematisches und alphabetisches Verzeichnis sowie hier insbesondere die Ergänzungsliste). Einzelne falsch eingeordnete oder mehrdeutige Diagnosentexte sollten dabei aussortiert, vor allem aber Daten aus der Praxis regelmäßig auf neu auftretende Umschreibungen überprüft und der Suchraum entsprechend

ergänzt werden. Trotz des umfangreichen Lernmaterials ist in der aktuellen Form eine Vielzahl an in der Klinik auftretenden Diagnosentexten nicht in den Lexika enthalten.

Der **Algorithmus** an sich bietet nur noch kleinere Verbesserungsmöglichkeiten; allerdings wurden gerade die ausgewerteten BFA-Diagnosen im Vorfeld dazu hergezogen, den Algorithmus zu verbessern, so dass dieses Ergebnis bei einem solchen Test, bei dem sich Lern- und Testdaten überlappen, auf der Hand liegt. Einige Vorschläge ergeben sich beim genaueren Betrachten aber auch hier. So sollte für die 4. Stelle zwischen den Konzepten „sonstige“ und „nicht näher bezeichnete“ der ICD unterschieden werden. Eine solche Unterscheidung wird durch den Algorithmus bisher nicht getroffen; dabei müsste eine Diagnose als „nicht näher bezeichnet“ klassifiziert werden, wenn sie ohne Zusätze auftritt, als „sonstige“, wenn bestimmte ergänzende Begriffe vorkommen, die entweder für jede ICD-Klasse definiert sein müssten (es gibt entsprechende Feinaufteilungen der ICD) oder algorithmisch bestimmten semantischen Kategorien zuzuordnen wären, die die Wertung eines Begriffs als für die Klassifikation relevanten Zusatz zumindest wahrscheinlich erscheinen ließe.

Mit dem derzeitigen Algorithmus sind manche Fehler auch schon aus rein stochastischen Gründen nicht vermeidbar. So ist es unter anderem nicht immer vorteilhaft, alle Krankheitsbegriffe an den ersten Rang bei der Suche zu stellen. Allerdings liefert eine andere Anordnung um so mehr Probleme an anderer Stelle. Abhilfe schaffen kann allenfalls ein kompletteres Durchsuchen des Suchraumes an sich, das bei der verbesserten Leistungsfähigkeit der inzwischen in der Praxis verwendeten EDV-Systeme durchaus auch in Echtzeit realistisch ist.

Das Zuordnen von Oberbegriffen mittels der Hierarchie von SNOMED II, auf dem die Überlegenheit des Retrievals beruht, muss durch eine Taxonomie ergänzt oder ersetzt werden, da sie bislang sehr unvollständig und zum Teil ungenau ist. Ein interessanter Ansatz hierfür findet sich in [Bousquet 00].

Zu erwähnen sei am Rande die Beobachtung, dass in einzelnen Testdaten (vgl. Tabelle 6.2.2) TNM-Klassifikationen auftreten, was ein Erkennen dieser Ausdrücke, mit denen Stadien maligner Prozesse klassifiziert werden, durch den Algorithmus als wünschenswert erscheinen lässt. –

Die übrigen Fehler, die den Algorithmus betreffen, verteilen sich auf alle seine Teile, und können nur punktuell und ohne große Effekte angegangen werden. Dennoch sollten auch

kleine Verbesserungen in Betracht gezogen werden, da einerseits der Aufwand bei Änderungen an der Datenbasis ungleich höher ist, und andererseits der Algorithmus im Gegensatz zu den zu Grunde liegenden Daten eine gewisse überragende Gültigkeit besitzt, also auf spätere Probleme mit veränderter Datenbasis oder andere Sprachen übertragbar ist. So konnte etwa der Teilalgorithmus zur morphologischen Zerlegung in [Hahn 01; Honeck 02] aufgegriffen und weiterentwickelt werden. Das Redigieren an vielen kleinen Stellen macht das System MedSearch zum Teil allerdings schon jetzt vergleichsweise komplex, und bereits kleine Verbesserungen, die meist um so unspektakulärere Wirkungen erzielen je subtiler sie sind, verursachen einen hohen und in der Praxis oft nicht vertretbaren Aufwand.

7.2 Leistungsfähigkeit von MedSearch im Vergleich

Die Resultate der drei in Abschnitt 6.2 präsentierten Verfahren weisen in absoluten Zahlen ein relativ niedriges Niveau an Übereinstimmung von *gold standard* und durch die einzelnen Ansätze ermittelten ICD-Klassen auf. Zwei Ursachen dafür sind in der mangelnden Qualität der manuellen Codierung des Arztes, die als *gold standard* verwendet wurde, sowie in der Kumulation identischer Texte zu einem einzigen Eintrag innerhalb den Testdaten zu suchen. Darüber hinaus wird mit allen drei Methoden bei der Verwendung von Kurzdiagnosen, wie sie einem automatischen Klassifikationssystem manuell eingegeben werden, eine ungleich höhere Quote an korrekten Ergebnissen erzielt, da ein entscheidender Teil der Fehlerquellen (zusätzliche Informationen, spezifische Abkürzungen) dann vom Arzt bewusst vermieden werden. Dennoch ist die Qualität der automatischen Codierung bei keiner der drei Methoden, die Freitextdiagnosen nach ICD-9 verschlüsseln, wirklich zufriedenstellend, und der Prozentsatz an korrekten Ergebnissen für ICD-10 und OPS-301-Klassifikation ist in einer ähnlichen Größenordnung zu erwarten.

Im Vergleich der einzelnen Verfahren untereinander ist zunächst eine geringfügige Verschlechterung vom lexikonfreien Ansatz (Trigramm-Indexierung + Vektorraum-Retrieval) zum wissensbasierten hin (SNOMED-Indexierung +Vektorraum-Retrieval) bemerkenswert. Die SNOMED-Indexierung an sich führt ohne adäquaten Gebrauch der durch die Indizes repräsentierten Information offensichtlich zu einem Qualitätsverlust, der durch die Komplexität der Indexierung durch SNOMED-Identifier gegenüber der Indexierung durch Trigramme verursacht wird. Diese Feststellung wurde bereits in [Brigl 95] getroffen, allerdings auf Basis einer unzureichend aussagekräftigen Datenmenge (17 Diagnosen). Die vergleichsweise ak-

zeptable Qualität des Retrievals auf Trigrammbasis resultiert dabei hauptsächlich aus dem beachtlichen Umfang des Suchraums, und auch die Leistungsfähigkeit anderer Ansätze korreliert trotz linguistisch motivierter Methoden in starkem Maße mit der Dichte des vorgegebenen ICD-Dokumentenraums.

Keine Erwähnung fand in der Untersuchung die Möglichkeit, bereits die morphologische Analyse als Mittel zur Indexierung von Texten durch Morpheme zu betrachten, an die sich dann bequem ein Vektorraum-Retrieval anschließen lässt. Dieses Verfahren liegt von den Resultaten her ziemlich genau in der Mitte zwischen dem auf Trigrammen und dem auf SNOMED-Indizes basierenden Vektorraum-Retrieval, so dass sich der Qualitätsverlust vom klassischen zum hybriden Ansatz zu ungefähr gleichen Maßen auf morphologische Analyse und SNOMED-Verschlüsselung aufteilt.

Signifikant besser als die vorangehenden Verfahren ist hingegen der aufwendigere, wissensbasierte MedSearch-Ansatz. Dafür ist, wie die Untersuchung der ersten beiden Methoden zeigt, nicht das Indexierungsverfahren ursächlich, sondern vielmehr das hier angewandte spezielle Retrieval, das die interne Struktur von SNOMED und die darin enthaltenen Informationen sinnvoll in den Algorithmus integriert und verwertet.

Neuere Referenznomenklaturen wie SNOMED RT und SNOMED CT [Spackman 98] beinhalten weitergehende Hierarchien und explizite Verweise auf ICD-Codes. Diese Hierarchien sind von der Notation unabhängig und in gesonderten Verzeichnissen hinterlegt [Brown 99]. Das MedSearch-Retrieval würde im Gegensatz zu anderen Ansätzen diese Informationen sinnvoll anwenden, und dadurch die Qualität der Codierung gegebenenfalls noch verbessern.

Erwähnung finden sollten zuletzt die beachtlichen Unterschiede im Antwort-Zeit-Verhalten der drei Ansätze. Besonders die klassische Methode fällt hier durch immensen Zeitverbrauch auf, der mit einer praktischen Umsetzung des Verfahrens nicht vereinbar ist. Der deutliche Unterschied zur zweiten Methode liegt darin begründet, dass die Rechenzeit für das Vektorraum-Retrieval mit Anzahl der Indizes pro Diagnose sehr stark ansteigt; bei der Trigrammindexierung wird jede Zeichenkette der Länge n jedoch durch n-2 Indizes repräsentiert, im Gegensatz zur SNOMED-Notation, bei der in der Regel nicht mehr als 2-5 Indexterme vorliegen. – Performanter als beide Methoden ist auch hier der MedSearch-Algorithmus, der im Test mit einem Prozessor der unteren Klasse (Pentium 2 133 MHz) bereits Antworten in Echtzeit ermöglicht, so dass unter diesem Aspekt sowohl dem Einsatz des Systems bei der Analyse und

Klassifikation großer Datensätze als auch der ferneren Entwicklung des Verfahrens bei Verwendung aktueller, leistungsstarker Hardware ein weiter Spielraum zur Verfügung steht.

7.3 Schlussfolgerungen und neue Anregungen

Mit MedSearch wurde ein Algorithmus zur automatischen Klassifikation von Diagnosen und Prozeduren entwickelt, der die Vorteile der SNOMED-Indexierung von Dokumenten mit einer neuartigen Retrievalmethode kombiniert.

Die Vorteile der *morphologischen Analyse* sind dabei längst anerkannt. Zu ihnen gehören unter anderem [Baud 98b, Baud 99] die Möglichkeit der quasi endlosen Wortakzeptanz bei vergleichsweise kleinem zu Grunde liegendem Morphemlexikon (vgl. Abschnitt 5.5), die technische Realisierbarkeit einer weitgehend korrekten Dekomposition und der Nutzen für eine weiterführende feinere semantische Analyse. Nachteile wie die zunehmend komplexer werdenden Regeln für das *Parsing* werden durch diese Vorzüge deutlich aufgewogen.

Die *SNOMED-Indexierung* an sich ist als Zwischenschritt auch für viele neuere und interessante Einsatzgebiete von Attraktivität, da mit zunehmender technischer Leistungsfähigkeit von EDV-Systemen sowie mit steigender Performanz der Indexieralgorithmen der Schritt weg von der Indexierung kurzer Diagnosen hin zu längeren Dokumenten längst getan wurde. Anwendungsmöglichkeiten wie Internetsuchverfahren sind zunehmend im Zentrum der Forschung, und Ansätze zur besseren semantischen Bearbeitung, zum Erkennen von synonymen Begriffen auch über Sprachbarrieren hinweg sowie zur besseren Beurteilung der wissenschaftlichen Relevanz eines Dokuments [Eysenbach 99], wie sie über die morphologische Analyse und SNOMED-Indexierung eines Dokuments existieren, werden mit Interesse verfolgt. Eine entscheidende Rolle kommt dabei der Qualität der SNOMED-Indexierung zu, die noch optimiert werden kann (in [Brigl 95] etwa werden rund 50 % von 385 Arztbriefdiagnosen vollständig korrekt SNOMED-indexiert, die anderen nur partiell); einige Ideen zur Weiterentwicklung existierender Verfahren wurden in der vorliegenden Arbeit präsentiert.

Der neuartige MedSearch-Algorithmus zum Retrieval und zur *ICD-Klassifikation* ist in vieler Hinsicht ein heuristischer, aber origineller Ansatz, der sicher noch weiterentwickelt und erforscht werden muss, um zu einer abschließenden Beurteilung gelangen zu können. Interessante Arbeiten in vergleichbarer Richtung gibt es indes bereits von anderen Arbeitsgruppen,

wie etwa [Bousquet 00], wo ein ähnlicher Ansatz verfolgt wird. Dabei stellt es sich als sinnvoll heraus, die SNOMED-Wissensbasis für die Definition einer semantischen Distanz heranzuziehen, ein Konzept das die beiden hier vorgestellten Ideen der Kategoriebildung (Abschnitt 4.3) und Hierarchisierung (Abschnitt 4.5) vereint und in einen allgemeingültigeren Kontext stellt. (Allerdings erfolgt in [Bousquet 00] das Retrieval nur auf Basis eines Teilbereichs der ICD, und die Zeiten für eine Klassifikation betragen mehrere Minuten.) Die Verwendung neuerer Terminologien wie SNOMED RT und SNOMED CT [Spackman 98] eröffnet den Zugang zu vorgefertigten multiplen Hierarchien und expliziten Abbildungen in die ICD. Letztere könnten in der Konsequenz dahingehend entwickelt werden, dass die eigentliche ICD-Klassifikation für den Benutzer unsichtbar bliebe, der anstelle fester Klassenbezeichnungen eine verständliche medizinische Terminologie verwenden würde, was sowohl dem Bedürfnis des Arztes nach einer exakten und medizinisch relevanten Nomenklatur als auch den Anforderungen der Verwaltung nach einer festen Menge an Diagnosen, um homogene Patientengruppen für Abrechnungs- und statistische Zwecke zu schaffen, entgegenkäme [Franz 00].

Als Wissensbasis stehen dem MedSearch-Ansatz umfangreiche *Lexika* zur Verfügung, deren Erstellung und Pflege bereits beim Entwurf eines vergleichbaren Algorithmus berücksichtigt werden muss. Insbesondere zu beachten ist der große Unterschied im benötigten Zeit- und Rechenaufwand für das Indexieren des umfangreichen Dokumentenraums versus dem raschen und wenig zeit- und kostenintensiven Retrieval einer Suchanfrage. Vergleichbare Beobachtungen lassen sich bekanntermaßen beim Unterhalt von Internetsuchmaschinen machen, bei denen für die Indexierung zum Teil mehre tausend Rechner parallel zur Verfügung stehen müssen, wohingegen eine Suchanfrage in Sekundenbruchteilen beantwortet werden kann.

Ein interessanter Gedanke grundsätzlicher Natur könnte bei künftigen Weiterentwicklungen ebenfalls Eingang finden. Das System MedSearch, das auf verschiedensten Lexika basiert, ist sehr eng auf diese Datenbasis zugeschnitten, die sich indes im Lauf der Zeit wandeln wird. So wurde der Übergang von ICD-9 zu ICD-10 bereits vollzogen; für das verwendete SNOMED II bestehen längst bisher allerdings rein englischsprachige Nachfolgenomenklaturen. Die Möglichkeit der Weiterentwicklung solcher Ordnungssysteme sollte daher bereits beim Entwurf des Algorithmus berücksichtigt werden durch den Einbau einer Taxonomie, die die Semantik von Änderungen der Datenbasis wie das Hinzunehmen, Streichen, Aufspalten von Klassen usw. zumindest im kleinen Rahmen, etwa für das Aktualisieren von Versionen, beschreibt [Cimino 96b].

Hilfreich und eine Überlegung wert ist es auch, die Codierung von Diagnosen oder medizinischen Texten aus einer anderen Blickrichtung heraus zu betrachten als eine Übersetzung von einer Sprache in eine zweite; die Ansätze und Probleme gleichen sich in mancher Hinsicht, und können sich gegenseitig befruchten. Dabei ist ein 1:1-Übertragen von natürlicher Sprache in die Notation eines (medizinischen) Ordnungssystems sicher nur für den Indexiervorgang denkbar, per Definition indes nicht für die Klassifikation. Für das Ordnungssystem müssen dabei allerdings zunehmend semantische Strukturen in Form einer „Grammatik“ entwickelt werden, um dem Ziel der 1:1-Umsetzung von natürlicher in Codesprache nahezukommen [Sager 95].

Die wachsende Bedeutung der Semantik in Relation zur Syntax generell hat ihre Ursache einerseits in der Verfügbarkeit leistungsstärkerer EDV-Systeme, die für die wesentlich komplexere Erfassung der Semantik Voraussetzung sind, und andererseits in der zunehmenden Annäherung der Möglichkeiten rein syntaktischer Algorithmen an eine Grenze der Codierqualität, deren Überschreiten nur durch zusätzliche semantische Konstrukte möglich ist. Die künftige Entwicklung wird daher immer mehr den Weg von Daten zu Konzepten beschreiben, indem zunehmend Modelle für die miteinander verknüpften Bereiche der Darstellung und des Verstehens medizinischer Fakten entwickelt werden und Eingang in die computerlinguistische Verarbeitung von Textdokumenten finden [Rassinoux 98]. Die Verlagerung von der reinen Repräsentation medizinischen Wissens hin zum semantischen *Understanding* ist in vollem Gange.

Bei all den Vorschlägen, die im Rahmen dieser Dissertation entwickelt wurden, sollte grundsätzlich der an sich triviale Gedanke nicht außer gelassen werden, dass mit jeder Neuerung auch der Arbeitsaufwand für ihre praktische Umsetzung ansteigt. Bereits die gründliche Erstellung und Pflege der Wissensbasis erfordert mehr zeitlichen und personellen Einsatz, als es für diese Studie teilweise möglich war. Die Verknüpfung semantischer und syntaktischer Methoden im MedSearch-Algorithmus ist zukunftsweisend; ein automatisches Klassifikationssystem für Diagnosen *ist* möglich, ja in qualitativ hochwertiger Form möglich. Die Entwicklung indes ist zeitintensiv und nur im Team umzusetzen [Gundersen 96].

Zum Schluss dieser Arbeit ist eine Überlegung angebracht, die bei der Suche nach Lösungen für die Verarbeitung medizinischer Dokumente aus Sicht der Informatik allein in aller Regel außer acht bleibt: Dass selbst die besten Verfahren ihre Grenzen dort haben, wo Kommunika-

tion besser ist als Dokumentation [Coiera 00]. Nur wenn dieser Aspekt ausreichend Berücksichtigung findet ist gewährleistet, dass auch künftige Entwicklungen in dem spannenden, sich zwischen syntaktischen und semantischen Methoden im Umbruch befindenden Gebiet der automatischen Indexierung und Klassifikation von Diagnosen letztlich dem Patienten als eigentlichem Objektiv aller Bemühungen im medizinischen Bereich zu Gute kommen.

Zusammenfassung

Einführung: Die Entwicklung neuer Indexier- und Klassifikationsalgorithmen für die Medizin bietet ein großes und spannendes Arbeitsfeld. In drei Abschnitten wird erläutert, wie 1. veränderte gesetzliche Verpflichtungen bezüglich der Dokumentation von Diagnosen und Prozeduren, 2. die Qualität in ihrer Leistungsfähigkeit begrenzter aktuell klinisch eingesetzter Verfahren der Codierung und 3. die deutlich angestiegenen Möglichkeiten der Informatik und Linguistik, verbunden mit der zunehmenden Verfügbarkeit performanter EDV-Systeme auf Station, Forschung auf diesem Gebiet notwendig machen und diese Arbeit motivieren.

Methoden: Ausgehend von einem früheren Ansatz wurde eine Methode zur Vorverarbeitung, morphologischen Analyse und SNOMED-Indexierung von Diagnosentexten umgesetzt und in vielen Punkten weiterentwickelt. Aufbauend auf dieser syntaktisch motivierten Indexierung schließt sich ein neuartiges Retrieval-Verfahren zur ICD- und OPS-301-Codierung an, das semantische Konzepte wie begriffliche Kategorien und eine Hierarchiebildung berücksichtigt. Eigenständig verwendet werden können sowohl die morphosyntaktische Zerlegung, die zum Teil bereits in neue Systeme integriert ist, als auch die SNOMED-Indexierung.

Realisation: Einige spezielle Aspekte der praktischen Umsetzung des beschriebenen Verfahrens werden für künftige Weiterentwicklungen näher erläutert, unter anderem zur Architektur und inneren Struktur des Algorithmus, zur Definition der Wissensbasis und zur Verwaltung des Morphemlexikons. Das Wachstumsverhalten eines Morphemlexikons in Relation zum Umfang des analysierbaren Wortschatzes wurde in einer eigenständigen Studie untersucht.

Ergebnisse: Zur Evaluation von Indexierung und Retrieval wurde ein qualitativer Test, der verbleibende Fehlerquellen aufdeckt und den Kategorien Testdaten, Lerndaten und Algorithmus zuordnet, sowie ein quantitatives Experiment zum Vergleich der Leistungsfähigkeit eines klassischen Ansatzes (Trigramm-Indexierung und Vektorraum-Retrieval) mit dem präsentierten Verfahren durchgeführt. Dabei ergab sich bei einer Anzahl von 2251 bzw. 10000 nach ICD-9 klassifizierten Testdatensätzen eine Übereinstimmung mit dem *gold standard* in 42,7 % bzw. 50,5 % der Fälle (Dreisteller).

Diskussion: Typische Fehlerquellen bei der automatischen Klassifikation sind die unzureichende Qualität der Testdaten, die Unvollständigkeit und mangelnde Aufarbeitung der Lerndaten sowie kleinere, systematische Fehler des Algorithmus. Für die Überlegenheit des präsentierten Verfahrens gegenüber dem konventionellen Ansatz ist nicht die Indexierungsmethode, sondern das Retrieval verantwortlich. Dies spiegelt die wachsende Bedeutung semantischer Konzepte gegenüber vornehmend syntaktischer Ansätze wieder und gibt dadurch die Grundrichtung für weitere Forschungsmöglichkeiten vor.

Anhang: Von MedSearch verwendete Tabellen

Um den Aufbau und die Realisation der MedSearch zu Grunde liegenden Wissensbasis zum detaillierteren Verständnis des Algorithmus sowie für Zwecke seiner möglichen Neuimplementation und Weiterentwicklung festzuhalten, werden im Folgenden die einzelnen Tabellen und ihr Inhalt kurz erläutert und anhand von Ausschnitten illustriert.

Insgesamt besteht diese Wissensbasis derzeit aus 18 Lexika, die hier im einzelnen aufgelistet und den Stufen des Algorithmus, für die sie benötigt werden, zugeordnet sind.

Vorverarbeitung:	(1) Alphabet
	(2) Months
	(3) Units
	(4) Numbers
	(5) Qualifiers
	(6) Abbreviations
Morphologische Zerlegung:	(7) Morphemes
Indexierung:	(8) SNOMED
Synonymverarbeitung:	(9) Synonymes
Klassifikation:	(10) ICD-9, ICD-10, OPS-301
	(11) ICD-9_ABC, ICD-10_ABC, OPS-301_ABC
	(12) Clues_ICD-9, Clues_ICD-10, Clues_OPS-301

(1) Alphabet

Diese Tabelle wird für eine erste Reduktion des als Grundmenge zugelassenen ASCII-Zeichensatzes verwendet. Sie umfasst zwei Spalten:

1) *ASCII* (Byte): Das zu ersetzende Zeichen. Indexiert (ohne Duplikate).
2) *NewASCII* (Byte): Das Zeichen, auf das abgebildet werden soll.

Ausschnitt aus der Tabelle „Alphabet“:

ASCII	**NewASCII**	
241	110	(ñ ⇨ n)
242	111	(ò ⇨ o)
243	111	(ó ⇨ o)
244	111	(ô ⇨ o)
245	111	(õ ⇨ o)
246	246	(ö ⇨ ö)
247	32	(÷ ⇨ Blank)
248	111	(ø ⇨ o)
249	117	(ù ⇨ u)
250	117	(ú ⇨ u)
251	117	(û ⇨ u)
252	252	(ü ⇨ ü)
253	121	(ý ⇨ y)

(2) Months

Die Tabelle „Months“ dient der Interpretation von Datumsangaben. Sie umfasst derzeit folgende Einträge:

Month	**Number**	**Abbreviation**
Jan	1	Ja
Januar	1	Nein
Feb	2	Ja
Febr	2	Ja
Februar	2	Nein
März	3	Nein
Apr	4	Ja
April	4	Nein
Mai	5	Nein
Jun	6	Ja
Juni	6	Nein
Jul	7	Ja
Juli	7	Nein
Aug	8	Ja
August	8	Nein
Sep	9	Ja
Sept	9	Ja
September	9	Nein
Okt	10	Ja
Oktober	10	Nein
Nov	11	Ja
November	11	Nein
Dez	12	Ja
Dezember	12	Nein

Die Spalte „Month“ ist indexiert (ohne Duplikate). Spalte „Number“ enthält die der Monatsbezeichnung zugeordnete Zahl, und Spalte „Abbreviation“ kennzeichnet abgekürzten Monatsnamen.

(3) Units

Die Tabelle „Units“ ist zur Umrechnung von Einheiten notwendig. Dabei wird eine Einheit (Spalte „Unit“) durch Multiplikation mit einem Faktor und nachfolgender Addition eines Summanden umgerechnet in die zugehörige Grundeinheit („BaseUnit“). Derzeit enthält diese Liste folgende Einträge:

Unit	**BaseUnit**	**Factor**	**Summand**
%		0,01	0
C	C	1	0
K	C	1	-273,15
F	C	0,5555556	-32
kg	g	1000	0
g	g	1	0
mg	g	0,001	0
µg	g	0,000001	0
Hz	Hz	1	0
kHz	Hz	1000	0
MHz	Hz	1000000	0
J	J	1	0
kJ	J	1000	0
cal	J	4,1868	0
kcal	J	4186,8	0
l	l	1	0
dl	l	0,1	0
cl	l	0,01	0
ml	l	0,001	0
µl	l	0,000001	0
km	m	1000	0
m	m	1	0
dm	m	0,1	0
cm	m	0,01	0
mm	m	0,001	0
µm	m	0,000001	0
nm	m	0,0000000001	0
Pa	Pa	1	0
mbar	Pa	100	0
bar	Pa	100000	0
ms	s	0,001	0
s	s	1	0

Minute	s	60	0
Minuten	s	60	0
min	s	60	0
Stunde	s	3600	0
Stunden	s	3600	0
h	s	3600	0
Tag	s	86400	0
Tage	s	86400	0
Tagen	s	86400	0
Woche	s	604800	0
Wochen	s	604800	0
Monat	s	2629800	0
Monate	s	2629800	0
Monaten	s	2629800	0
Jahr	s	31557600	0
Jahre	s	31557600	0
Jahren	s	31557600	0

Spalte „Unit“ ist indexiert (ohne Duplikate).

(4) Numbers

Die Tabelle „Numbers“ dient der Abbildung römischer Zahlen in arabische und umfasst derzeit folgende Einträge:

Roman	**Arabic**				
I	1	V	5	IXc	9c
Ia	1a	Va	5a	X	10
Ib	1b	Vb	5b	XI	11
Ic	1c	Vc	5c	XII	12
Id	1d	VI	6	XIII	13
II	2	VIa	6a	XIV	14
IIa	2a	VIb	6b	XIX	19
IIb	2b	VIc	6c	XV	15
IIc	2c	VII	7	XVI	16
IId	2d	VIIa	7a	XVII	17
III	3	VIIb	7b	XVIII	18
IIIa	3a	VIIc	7c	IXX	19
IIIb	3b	VIII	8	XX	20
IIIc	3c	VIIIa	8a	XXI	21
IIId	3d	VIIIb	8b	XXII	22
IV	4	VIIIc	8c	XXIII	23
IVa	4a	IX	9	XXIV	24
IVb	4b	IXa	9a	XXV	25
IVc	4c	IXb	9b		

Die Einträge der Spalte „Roman“ sind indexiert (ohne Duplikate) und werden durch den Algorithmus auf die Einträge der Spalte „Arabic“ abgebildet.

(5) Qualifiers

Die Tabelle „Qualifiers“ umfasst derzeit folgende Informationsqualifikatoren (nach Brigl):

Entry	**Qualifier**
Ausschluss	nicht
nicht	nicht
DD	Differenzialdiagnose
n.n.bez	nicht näher bezeichnet
n.n.bez.	nicht näher bezeichnet
n.n.einzuordnende	nicht näher einzuordnend
V.a.	Verdacht auf
V. a.	Verdacht auf
Verdacht auf	Verdacht auf
Z.bei	Zustand bei
Z. bei	Zustand bei
Zustand bei	Zustand bei
Z.n.	Zustand nach
Z. n.	Zustand nach
Zust.n.	Zustand nach
Zust. n.	Zustand nach
Zust. nach	Zustand nach
Zustand nach	Zustand nach

Die Spalte „Entry“ ist indexiert (ohne Duplikate), ihre Einträge werden durch die der Spalte „Qualifier“ ersetzt und als Informationsqualifikator gekennzeichnet.

(6) Abbreviations

Diese Tabelle dient der Auflösung von Abkürzungen. Sie enthält vier Spalten:

1) *Identifier* (Zähler): Indexiert (ohne Duplikate).
2) *Abbreviation* (String der Länge 30): Der abgekürzte Text. Indexiert (Duplikate möglich).
3) *PlainText* (String der Länge 100): Volltextfassung, durch die ersetzt wird.
4) *Type* (Integer): 1 = abgekürztes Wort mit Punkt am Ende; 2 = abgekürztes Wort in Großbuchstaben; 3 = abgekürztes Wortende mit Punkt am Ende (z.B. *krkht. ⇨ *krankheit).

Ausschnitt aus der Tabelle „Abbreviations“:

Identifier	Abbreviation	PlainText	Type
327	kollat.	kollateral	1
328	Kompl.	Komplikation	1
329	Komplik.	Komplikation	1
330	kongenit.	kongenital	1
331	konst.	konstant	1
332	kontralat.	kontralateral	1
333	KPL	Keratoplastik	2
334	kran.	kranial	1
335	krkh.	krankheit	3
336	krkht.	krankheit	3
337	Krkht.	Krankheit	1
338	Krkhtt.	Krankheiten	1

(7) Morphemes

Diese Tabelle enthält das Morphemlexikon. Sie unterteilt sich in 6 Spalten:

1) *Identifier* (Zähler): Indexiert (Ohne Duplikate).
2) *Original* (String): Das Morphem vor der Vorverarbeitung.
3) *Morpheme* (String): Das Morphem nach der Vorverarbeitung. Indexiert (Duplikate möglich).
4) *Example* (String): Ein Beispiel für das Vorkommen des Morphems.
5) *Type* (Integer): Der Morphemtyp.
6) *Weight* (Integer: Das Gewicht des Lexems (0, 1 oder 2).

Die *Originalzeichenkette* (Spalte 2) sollte unbedingt gespeichert werden, damit sie bei Veränderungen der Vorverarbeitung noch zur Verfügung steht.

Ein *Beispiel* (Spalte 4) dient dazu, den Ursprung vor allem kleiner, wenig signifikanter Morpheme festzuhalten, um bei einer Revision des Morphemlexikons den Grund für die Aufnahme eines Morphems nachvollziehen zu können.

Der Typ (Spalte 5) eines Morphems wird derzeit gekennzeichnet durch eine Zahl zwischen 1 und 10. Dabei stehen die einzelnen Werte für:

1 = Wortstamm, 2 = Präfix, 3 = Derivationssuffix, 4 = fester Begriff, 5 = Infix, 6 = Flexionssuffix, 7 = automatisch erzeugtes Morphem, 8 = Zahl, 9 = automatisch erzeugter Suffix, 10 = automatisch erzeugter Präfix.

Infixe sind dabei lediglich die drei Morpheme s, co und o. Als *Flexionssuffixe* fungieren aktuell s, es, e, em, en, er und o. Da sie größtenteils auch als Derivationssuffixe in Erscheinung treten, sind sie von nachrangiger Bedeutung. Typ 7 bis 10 werden derzeit wie Typ 4 (fester Begriff) behandelt.

Ausschnitt aus der Tabelle „Morphemes“:

Identifier	**Original**	**Morpheme**	**Example**	**Type**	**Weight**
291	gesicht	gesicht	Gesichtsschädelfrakturen	1	2
292	schädel	schaedel	Gesichtsschädelfrakturen	1	2
293	neben	neben	Nasennebenhöhlenpunktion	2	2
294	punkt	punkt	Nasennebenhöhlenpunktion	1	2
295	endo	endo	Kiefergelenkendoprothese	2	2
296	einstell	einstell	Einstellungsuntersuchung	1	2
297	ix	ix	Zervixstumpfexstirpation	3	1

(8) SNOMED

Diese Tabelle umfasst die Nomenklatur SNOMED II, mit einigen für den Algorithmus wichtigen Ergänzungen. Sie enthält 7 Spalten:

1) *Identifier* (Zähler): Indexiert (ohne Duplikate).
2) *Morpheme* (Long Integer): Der Identifier eines Morphems, das in der Zerlegung des Eintrags (Spalte 7) vorkommt. Indexiert (Duplikate möglich).
3) *Code* (String der Länge 6): Der SNOMED-Code. Indexiert (Duplikate möglich).
4) *Order* (Character): Leer für Haupteintrag, S für Synonym, R für verwandter Begriff, T für veralteter Begriff.
5) *Reference* (String der Länge 30): Bei Einträge der Kategorie „verwandter Begriff“ werden hier oft weitere SNOMED-Codes angegeben, die z.B. einen topographischen Sitz bezeichnen können.
6) *Entry* (String der Länge 120): Der SNOMED-Eintrag.
7) *Parsation* (String der Länge 150): Die Zerlegung des Eintrags (d.h. die Morphemidentifier in der Reihenfolge der Zerlegung; Wortgrenzen werden durch Klammern bezeichnet).

Ausschnitt aus der Tabelle „SNOMED“:

Identif.	Mor.	Code	O.	Ref.	Entry	Parsation
102480	2256	D12000			Krkht. des Purinmetabolismus	(66 5) (2256 679 108)
102481	4927	D12000	R		Krkht. des Nukleinsäuremetabolismus	(66 5) (4927 2114 846 679 108)
102482	895	D12000	R		Krkht. des Pyrimidinmetabolismus	(66 5) (895 5418 679 108)
102483	3301	D12010			prim. Gicht-Syndrom	(3756 486) (3301) (72)
102484	6129	D12010	R	T12000	Arthritis divitum	(364 135) (6129)
102485	4451	D12010	R	T12000	Arthritis urica	(364 135) (4451)
102486	3568	D12010	S		Diathesis urica	(3568 754) (4451)
102487	3301	D12010	S		Gicht	(3301)
102488	3301	D12010	R	T12000	Gichtarthritis	(3301 364 135)
102489	145	D12010	S		Harnsäuregicht	(145 846 3301)
102490	6130	D12010	S		Hyperurikopathie-Syndrom	(6130) (72)
102491	6131	D12010	S		Morbus aulicus	(4221 150) (6131)
102492	2520	D12010	S		prim. fam. Hyperurikämie	(3756 486) (1156 486) (78 2520)
102493	6132	D12010	S		uratische Diathese	(6132) (3568)
102494	6133	D12010	S		Urikopathie-Syndrom	(6133) (72)
102495	4476	D12040			LESCH-NYHAN Syndrom	(4476) (4477) (72)
102496	6135	D12040	R		HG-PRT-Mangel-Syndrom	(109 2124) (6134) (6135) (2174 5098) (105) (72)
102497	2520	D12040	S		Hyperurikämie-Syndrom	(78 2520) (72)
102498	6136	D12040	R		Hypoxanthin-guanin-phosphoryl-transferase-Mangel-Syndrom	(109 2124) (6134) (6136) (2174 5098) (105) (72)
102499	26240	D12050			Orotazidurie-Syndrom	(26240) (72)
102500	2124	D12060			Xanthinurie	(2124 2500)
102501	15	D12500			Störung des Kohlenhydratstoffwechsels	(15 8) (64 1 194 51 122)
...						
166257	1181	T12000			Articulatio	(1181 51 363)

Dabei entsprechen die numerischen Identifier folgenden Morphemen:

(66 5) (2256 679 108) = (krank-heit) (purin-metabol-ismus)

(66 5) (4927 2114 846 679 108) = (krank-heit) (nukl-ein-saeure-metabol-ismus)

(66 5) (895 5418 679 108) = (krank-heit) (pyrimid-in-metabol-ismus)

(3756 486) (3301) (72) = (prim-aer) (gicht) (syndrom)

usw.

(9) Synonymes

Diese Tabelle enthält synonyme Begriffe, die nicht durch SNOMED bereits abgedeckt sind. Sie umfasst folgende 5 Spalten:

1) *Code* (String der Länge 6): Ein SNOMED-Code, der in der Index1 vorkommt. Indexiert (Duplikate möglich).
2) *Entry1* (String der Länge 100): Das zu ersetzende Wort.
3) *Entry2* (String der Länge 100): Das Synonym für Entry1.
4) *Index1* (String der Länge 100): Die SNOMED-Indexierung (inkl. L-Achse) von Entry1.
5) *Index2* (String der Länge 100): Die SNOMED-Indexierung (inkl. L-Achse) von Entry2.

Ausschnitt aus der Tabelle „Synonymes“:

Code	**Entry1**	**Entry2**	**Index1**	**Index2**
L00713	rekonstruktive Chirurgie	Wiederherstellungs-chirurgie	L00713 PY9003	P14700
L00718	Verkehrsweg	Fahrweg	L00718 TYX466	E91560 TYX466
L00725	vorsätzlich	absichtlich	L00725	L00374
L00739	Straße	Fahrweg	L00739	E91560 TYX466
L00772	beidseitig	bilateral	L00772	T00320
L00874	verlängertes Koma	Coma prolongé	L00874 F85640	F85740

(10) ICD-9/ICD-10/OPS-301

Diese drei Tabellen enthalten die systematischen Verzeichnisse von ICD-9, ICD-10 und OPS-301 in einer Volltextfassung. Sie sind identisch aufgebaut und umfassen folgende 4 Spalten:

1) *Identifier* (Zähler): Indexiert (ohne Duplikate).
2) *Code* (String der Länge 7): Die Notation der ICD- bzw. OPS-301-Klasse. Indexiert (Duplikate möglich).
3) *Order* (Character): Ebene des Eintrags. Dabei steht # für Kapitel, $ für Abschnitt, & für Unterabschnitt, < und > für dreistellige (ICD) / vierstellige (OPS-301) Notationen (wobei < nicht in Viersteller / Fünfsteller unterteilt ist), § für vierstellige Notation (OPS-301: fünfstellig).
4) *Entry* (String der Länge 240): Der Eintrag in der Volltextfassung.

Da durch den MedSearch-Algorithmus stets eine Notation mit maximaler Anzahl an Stellen ermittelt wird, sind hier nur Codes der Ordnung < und § relevant.

Ausschnitt aus der Tabelle „ICD-9“:

Identifier	**Code**	**Order**	**Entry**
1565	271.-	>	Störungen des Kohlenhydrattransportes und –stoffwechsels
1566	271.0	§	Glykogenosen
1567	271.1	§	Galaktosämie
1568	271.2	§	Hereditäre Fruktoseintoleranz
1569	271.3	§	Disaccharidintoleranzsyndrom und Glukose-Galaktose-Malabsorption
1570	271.4	§	Renale Glykosurie
1571	271.8	§	Sonstige Störungen des Kohlenhydrattransportes und -stoffwechsels
1572	271.9	§	Nicht näher bezeichnete Störung des Kohlenhydrattransportes und -stoffwechsels

(11) ICD-9_ABC/ICD-10_ABC/OPS-301_ABC

Diese drei Tabellen, die identisch aufgebaut sind, enthalten das systematische, alphabetische und Ergänzungsverzeichnis der ICD-9, ICD-10 bzw. des OPS-301 in Originalform sowie deren SNOMED-Indexierung, wie sie mit dem Algorithmus ermittelt wurde. Sie umfassen je 5 Spalten:

1) *Identifier* (Zähler): Indexiert (ohne Duplikate). Dieser Bezeichner wird von der Tabelle Clues_ICD-9 (resp. Clues_ICD-10 bzw. Clues_OPS-301) verwendet.
2) *Code* (ICD: String der Länge 5; OPS-301: String der Länge 8): Die ICD- bzw. OPS-301-Notation. Indexiert (Duplikate möglich).
3) *SNOMED* (String der Länge 255): Die Zerlegung des Originaleintrags mittels SNOMED (inklusive L-Achse).
4) *Entry* (String der Länge 255): Der Originaleintrag.
5) *Source* (String der Länge 3): Ursprung des Eintrags (ICD = Systematik, ABC = Alphabet, ERG = Ergänzungen).

Ausschnitt aus der Tabelle „ICD-9_ABC“:

Identifier	Code	SNOMED	Entry	Source
15964	271.3	D12800^D62080	Disaccharidintoleranzsyndrom und Glukose-Galaktose- Malabsorption	ICD
15965	271.3	L02114^DY0610^ F01880	Laktoseintoleranzsyndrom	ERG
15966	271.3	D12800	Disaccharid Intoleranz	ABC
15967	271.3	D12800	Disaccharid Malabsorption	ABC
15968	271.3	D12800	Intoleranz Disaccharid	ABC
15969	271.3	F01920^F11820^ F11900	Intoleranz Glucose Galaktose	ABC
15970	271.3	D12840	Intoleranz Laktose	ABC
15971	271.3	L28082^F01920^ F11980	Intoleranz Saccharose Isomaltose	ABC
15972	271.3	EY0019^F65930	Laktosurie	ABC
15973	271.3	D12800	Malabsorption Disaccharid	ABC
15974	271.3	F62002^F11820	Malabsorption Glukose	ABC
15975	271.3	L28082^F62002	Malabsorption Isomaltose	ABC
15976	271.3	D12840	Malabsorption Laktose	ABC
15977	271.3	L01790^L00999^ F62002	Malabsorption Monosaccharid	ABC
15978	271.3	F62002^F11980	Malabsorption Saccharose	ABC
15979	271.3	FYX101^F21100	Mangel Disaccharidase	ABC

Die Einträge aus dem Alphabet wurden um alle fakultativen Ergänzungen (durch Klammerung zu erkennen) gekürzt; bestimmte Floskeln wurden gestrichen.

Alphabet-Einträge im Original:

271.3 Disaccharid | Intoleranz (hereditäre)
271.3 Disaccharid | Malabsorption
271.3 Intoleranz | Disaccharid- (hereditäre)
271.3 Intoleranz | Glucose-Galaktose
271.3 Intoleranz | Laktose- (hereditäre) (infantile)
271.3 Intoleranz | Saccharose-Isomaltose
271.3 Laktosurie
271.3 Malabsorption | Disaccharid
271.3 Malabsorption | Glukose (-Galaktose)
271.3 Malabsorption | Isomaltose
271.3 Malabsorption | Laktose (hereditäre)
271.3 Malabsorption | Monosaccharid
271.3 Malabsorption | Saccharose (-Isomaltose)
271.3 Mangel (an) \ Mangelkrankheit (durch) | Disaccharidase

Im letzten Eintrag wurde z.B. auch die Alternative \...| entfernt. Generell sollten geklammerte Bestandteile nicht unabdingbar sein, können ggf. aber zusätzliche Information enthalten; daher wäre ein Mitverwenden zu diskutieren. Bei einigen Codes war allerdings eine Nachkor-

rektur von Hand notwendig, insbesondere 042-044 AIDS/HIV in einigen Fällen der Nachtrag von „AIDS“ und bei malignen Neubildungen 140-208 z.T. der Nachtrag „bösartig“.

(12) Clues_ICD-9/Clues_ICD-10/Clues_OPS-301

Diese Tabellen enthalten die Verweise von SNOMED-Codes auf Listen von ICD- bzw. OPS-301-Einträgen, in deren Indexierung sie enthalten sind. Dabei wird jeder Eintrag der Tabellen ICD-9_ABC, ICD-10_ABC bzw. OPS-301_ABC genau einmal repräsentiert. Sie enthalten folgende drei Spalten:

1) *Code* (String der Länge 6): Ein SNOMED-Code. Indexiert (Duplikate möglich).
2) *Source* (String der Länge 250): Eine Liste von Bezeichnern, die auf einige (nicht alle!) Einträge in der Tabelle ICD-9_ABC (resp. ICD-10_ABC bzw. OPS-301_ABC) verweisen, in deren Indexierungen der SNOMED-Code vorkommt.
3) *Class* (String der Länge 250): Die zugehörigen ICD-Klassifikationscodes.

Die Beschränkung der Längen der Strings (Zeichenketten) bringt mit sich, dass überlange Strings auf mehrere Zeilen in der Tabelle verteilt werden müssen. Deshalb sind Duplikate im Feld „Code“ erlaubt.

Ausschnitt aus der Tabelle „Clues_ICD-9“:

Code	**Source**	**Class**
D12010	4545^16205^16206^16210 ^16213 ^...	095^274.0^274.0^274.0^274.0 ^...
D12040	16471^	277.2^
D12060	16473^	277.2^
D12500	15985^15992^15993^	271.8^271.9^271.9^
D12510	15892^15902^15914^15915^ 15916^...	271.0^271.0^271.0^271.0^271.0^...
D12520	15904^15905^15917^	271.0^271.0^271.0^
D12530	15894^15921^	271.0^271.0^
D12540	15903^	271.0^
D12550	15893^16437^	271.0^277.0^
D12571	15908^	271.0^
D12572	15938^	271.0^

Literatur

[Baud 98a] Baud RH, Lovis C, Rassinoux AM, Scherrer JR (1998). *Alternative Ways for Knowledge Collection, Indexing and Robust Language Retrieval.* Methods of Information in Medicine 1998; 37: 315-326

[Baud 98b] Baud RH, Lovis C, Rassinoux AM, Scherrer JR (1998). *Morpho-Semantic Parsing of Medical Expressions.* Proceedings of the 1998 AMIA Annual Fall Symposium, 760-764

[Baud 99] Baud RH, Rassinoux AM, Rush P, Lovis C, Scherrer JR (1999). *The power and limits of a rule-based morpho-syntactic parser*. Proceedings of the 1999 AMIA Annual Fall Symposium, 22-26

[Blanquet 99] Blanquet A, Zweigenbaum P. *A Lexical Method for Assisted Extraction and Coding of ICD-10 Diagnoses from Free Text Patient Discharge Summaries.* Proceedings of the 1999 AMIA Annual Fall Symposium, 1029

[Bousquet 00] Bousquet C, Jaulent MC, Chatellier G, Degoulet P (2000). *Using Semantic Distance for the Efficient Coding of Medical Concepts.* Proceedings of the 2000 AMIA Annual Fall Symposium, 96-100

[Bosing 96] Bosing-Schwenkglenks M, Swoboda A, Blumenstock G (1996). *Qualität der Diagnosecodierung am Beispiel von drei ausgewählten Fachabteilungen des Universitätsklinikums Tübingen.* f & w 6/96; 13. Jahrgang: 593-597

[Brigl 94] Brigl B, Mieth M, Haux R, Glück E (1994). *The LBI-method for automated indexing of diagnoses by using SNOMED. Part I. Design and realization.* International Journal of Biomedical Computing 37: 237-247

[Brigl 95] Brigl B, Mieth M, Haux R, Glück E (1995). *The LBI-method for automated indexing of diagnoses by using SNOMED. Part II. Evaluation.* International Journal of Biomedical Computing 38: 101-108

[Brown 99] Brown PJB, Price C (1999). *Semantic Based Concept Differential Retrieval & Equivalence Detection in Clinical Terms Version 3 (Read Codes).* Proceedings of the 1999 AMIA Annual Fall Symposium, 27-31

[Bruijn 96] de Bruijn LM, Verheijen E, van Nes FL, Arends JW (1996). *Assigning SNOMED codes to natural language pathology reports.* Medical Informatics Europe 1996, 198-202

[Bruijn 97] de Bruijn LM, Hasman A, Arends JW (1997). *Automatic SNOMED classification – a corpus-based method.* Computer Methods and Programs in Biomedicine 54: 115-122

[Bruijn 98] de Bruijn LM, Hasman A, Arends JW (1998). *Automatic Coding of Diagnostic Reports.* Methods of Information in Medicine 1998; 35: 260-265

[Campbell 97] Campbell JR, Carpenter P, Sneiderman C, Cohn S, Chute CG, Warren J (1997). *Phase II Evaluation of Clinical Coding Schemes: Completeness, Taxonomy, Mapping, Definitions, and Clarity.* JAMIA 1997 (4): 238-251

[Carter 96] Carter KJ, Rinehart S, Kessler E, Caccamo LP, Ritchey NP, Erickson BA, Castro F, Poggione MD (1996). *Quality Assurance in Anatomic Pathology: Automatic SNOMED Coding.* JAMIA 1996 (3): 270-272

[Cimino 96a] Cimino JJ (1996). *Coding Systems in Health Care.* Review Paper. Methods of Information in Medicine 1996; 35: 237-284

[Cimino 96b] Cimino JJ (1996). *Formal Descriptions and Adaptive Mechanisms for Changes in Controlled Medical Vocabularies.* Methods of Information in Medicine 1996; 35: 202-210

[Cimino 98] Cimino JJ (1998). *Desiderata for Controlled Medical Vocabularies in the Twenty-First Century.* Methods of Information in Medicine 1998; 37: 394-403

[Coiera 00] Coiera E (2000). *When conversation is better than computation.* JAMIA 2000 (7): 277-286

[Diekmann 92] Diekmann F, Ehlers C-T, Eichhorn S, Kolodizig C (1992). *Diagnosenstatistik. Einsatz im Krankenhaus und für Pflegesatzverhandlungen.* Forschungsbericht des Bundesministeriums für Gesundheit. Nomos, Baden-Baden

[Dolin 01] Dolin RH, Spackman K, Abilla A, Correia CM, Goldberg B, Konicek D, Lukoff J, Lundberg C (2001). *The SNOMED RT Procedure Model.* Proceedings of the 2001 AMIA Annual Fall Symposium, 139-143

[Elkin 01] Elkin PL, Ruggieri AP, Brown S, Buntrock J, Bauer BA, Wahner-Roedler D, Litin SC, Beinborn J, Bailey KR, Bergstrom L (2001). *A Randomized Controlled Trial of the Accuracy of Clinical Record Retrieval using SNOMED-RT as Compared with ICD9-CM.* Proceedings of the 2001 AMIA Annual Fall Symposium, 159-163

[Espino 01] Espino JU, Wagner MM (2001). *Accuracy of ICD-9-coded Chief Complaints and Diagnoses for the Detection of Acute Respiratory Illness.* Proceedings of the 2001 AMIA Annual Fall Symposium, 164-168

[Eysenbach 99] Eysenbach G, Diepgen TL (1999). *Labeling and Filtering of Medical Information on the Internet.* Methods of Information in Medicine 1999; 38: 80-88

[Franz 00] Franz P, Zaiß A, Schulz S, Hahn U, Klar R (2000). *Automated Coding of Diagnoses – Three Methods Compared.* Proceedings of the 2000 AMIA Annual Fall Symposium, 250-254

[Gaus 00] Gaus W (2000). *Dokumentations- und Ordnungslehre.* 3. Aufl. Springer, Berlin, Heidelberg, New York.

[Gersenovic 95] Gersenovic M (1995). *The ICD Family of Classifications.* Methods of Information in Medicine 1995; 34: 172-175

[GMDS 96] (1996) *GMDS-Stellungnahme zur Klassifikation und Datenübermittlung von Diagnosen und Operationen.* Informatik, Biometrie und Epidemiologie in Medizin und Biologie 27 (2), Mitteilungen.

[GMDS 97] Zaiß A, Bülzebruck H, Glück E, Graubner B, Leiner F, Lochmann U, Straube R, Thurmayr R (1997). *Leitfaden zur medizinischen Basisdokumentation nach § 301 SGB V.* Arbeitsgruppe „Medizinische Dokumentation und Klassifikation“ der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS), Deutsche Krankenhaus Verlagsgesellschaft mbH

[Graubner 95] Graubner B (1995). *Wesentliche Klassifikationen für die medizinische Dokumentation in Deutschland und ihr Entwicklungsstand.* In: ICIDH – International Classification of Impairments, Disabilities and Handicaps. Ullstein Mosby, 41-69

[Gundersen 96] Gundersen ML, Haug PJ, Pryor TA, van Bree R, Koehler S, Bauer K, Clemons B (1996). *Development and Evaluation of a Computerized Admission Diagnoses System.* Computer and Biomedical Research 29: 351-372

[Hahn 01] Hahn U, Honeck M, Piotrowski M, Schulz S (2001). *Subword segmentation: Leveling out morphological variations for medical document retrieval.* Proceedings of the 2001 AMIA Annual Fall Symposium, 229-233

[Hersh 95] Hersh W, Leone TJ (1995). *The SAPHIRE Server: A New Algorithm and Implementation.* Proceedings of the 1995 AMIA Annual Fall Symposium, 858-862

[Hersh 96] Hersh W (1996). *Information Retrieval – A health care perspective.* Springer, New York

[Hohnloser 96] Hohnloser JH, Kadlec P, Pürner F (1996). *Coding Clinical Information: Analysis of Clinicians Using Computerized Coding.* Methods of Information in Medicine 1996; 35: 104-107

[Honeck 02] Honeck M, Hahn U, Klar R, Schulz S (2002). *Text Retrieval Based on Medical Subwords.* Medical Informatics Europe 2002, 241-245

[Ingenerf 97] Ingenerf J (1997). *Medizinische Linguistik.* In: Seelos HJ (Hrsg): *Medizinische Informatik, Biometrie und Epidemiologie.* De Gruyter, Berlin, 13-42

[Johnson 99] Johnson SB (1999). *A Semantic Lexicon for Medical Language Processing.* JAMIA 1999 (6): 205-218

[Kiuchi 95] Kiuchi T, Ohashi Y, Sato H, Kaihara S (1995). *Methodology for the Construction of a Disease Nomenclature and Classification System for Clinical Use.* Methods of Information in Medicine 1995; 34: 511-517

[Klar 97] Klar R, Graubner B (1997). *Medizinische Dokumentation.* In: Seelos HJ (Hrsg): *Medizinische Informatik, Biometrie und Epidemiologie.* De Gruyter, Berlin, 13-42

[Knüppel 01] Knüppel A (2001). *Untersuchungen zum Zipf-Mandelbrot-Gesetz an deutschen Texten.* In: Best KH (Hrsg): *Häufigkeitsverteilungen in Texten.* Peust & Gutschmidt, Göttingen, 248-280

[Lauterbach 00] Lauterbach K, Lüngen M (2000). *DRG-Fallpauschalen – Eine Einführung.* Schattauer, Stuttgart

[Lovis 00] Lovis C, Baud R (2000). *Fast Exact String Pattern-matching Algorithms Adapted to the Characteristics of the Medical Language.* JAMIA 2000 (7): 378-391

[Lussier 98] Lussier YA, Rothwell DJ, Côté RA (1998). *The SNOMED Model: A Knowledge Source for the Controlled Terminology of the Computerized Patient Record.* Methods of Information in Medicine 1998; 37: 161-164

[Lussier 01] Lussier YA, Shagina L, Friedman C (2001). *Automating SNOMED Coding Using Medical Language Understanding: A Feasibility Study.* Proceedings of the 2001 AMIA Annual Fall Symposium, 418-422

[Metzger 02] Metzger M, Köninger H (2002). *Anforderungen an das Berichtswesen im Zeitalter von DRGs.* In: *DRG-Einführung in deutschen Krankenhäusern.* Sonderbeilage „Das Krankenhaus“, Kohlhammer

[Michel 95] Michel PA, Lovis C, Baud R (1995). *LUCID: A Semi-automated ICD-9 Encoding System.* Medinfo 1995 – Proceedings of the 8th Conference on Medical Informatics, 1656

[Nietzschke 92] Nietzschke E, Wiegand M (1992). *Fehleranalyse bei der Diagnosenverschlüsselung nach ICD-9 gemäß der Bundespflegesatzverordnung.* Zeitschrift für Orthopädie 130: 382-387

[Nilsson 00] Nilsson G, Petersson H, Åhlfeldt H, Strender LE (2000). *Evaluation of Three Swedish ICD-10 Primary Care Versions: Reliability and Ease of Use in Diagnostic Coding.* Methods of Information in Medicine 2000; 39: 325-331

[O'Neil 95] O'Neil M, Payne C, Read J (1995). *Read Codes Version 3: A User Led Terminology.* Methods of Information in Medicine 1995; 34: 187-192

[Rechenberg 02] Rechenberg P, Pomberger G (Hg.) (2002). *Informatik-Handbuch.* 3. Aufl. Hanser, München

[Rector 99] Rector AL (1999). *Clinical Terminology – Why Is it so Hard?* Methods of Information in Medicine 1999; 38: 239-252

[Roeder 02] Roeder N, Irps S, Juhra C, Glockner S, Fiori W, Müller ML, Hecht A (2002). *Erlöse sichern durch Kodierqualität.* In: *DRG-Einführung in deutschen Krankenhäusern.* Sonderbeilage „Das Krankenhaus", Kohlhammer

[Rothwell 95] Rothwell DJ (1995). *SNOMED-Based Knowledge Representation.* Methods of Information in Medicine 1995; 34: 209-213

[Ruch 99] Ruch P, Wagner J, Bouillon P, Baud RH, Rassinoux AM, Scherrer JR (1999). *MEDTAG: Tag-like Semantics for Medical Document Indexing.* Proceedings of the 1999 AMIA Annual Fall Symposium, 137-141

[Sager 95] Sager N, Lyman M, Nhàn NT, Tick LJ (1995). *Medical Language Processing: Applications to Patient Data Representation and Automatic Encoding.* Methods of Information in Medicine 1995; 34: 140-146

[Salton 91] Salton G (1991). *Developments in Automatic Text Retrieval.* Science 253: 974-980

[Salton 94] Salton G, Allan J, Buckley C (1994). *Automatic Structuring and Retrieval of Large Text Files.* Communications of the ACM 37(2): 97-108

[Schlottmann 02] Schlottmann N, Raskop AM (2002). *Deutsche Kodierrichtlinien für Krankenhäuser*. In: *DRG-Einführung in deutschen Krankenhäusern.* Sonderbeilage „Das Krankenhaus“, Kohlhammer

[Schulz98a] Schulz S, Romacker M, Hahn U (1998). *Part-Whole Reasoning in Medical Ontologies Revisited – Introducing SEP Triplets into Classification-Based Description Logics.* Proceedings of the 1998 AMIA Annual Fall Symposium, 830-834

[Schulz 98b] Schulz S, Zaiß A, Brunner R, Spinner D, Klar R (1998). *Conversion Problems concerning Automated Mapping from ICD-10 to ICD-9.* Methods of Information in Medicine 1998; 37: 254-259

[Schulz 99] Schulz S, Romacker M, Franz P, Zaiß A, Klar R, Hahn U (1999). *Towards a multilingual morpheme thesaurus for medical free-text retrieval.* Medical Informatics Europe 1999, 891-894

[Schulz 00] Schulz S, Hahn U (2000). *Morpheme-Based, Cross-Lingual Indexing for medical Document Retrieval.* International Journal of Medical Informatics 2000; 58-59: 87-99

[Sozialgesetzbuch 96] (1996) *Krankenhausrecht.* Sozialgesetzbuch, Buch V, § 301, Taschenausgabe, 240 ff.

[Spackman 98] Spackman KA, Campbell KE (1998). *Compositional concept representation using SNOMED: towards further convergence of clinical terminologies.* Proceedings of the 1998 AMIA Annual Fall Symposium, 740-744

[Spyns 96] Spyns P (1996). *Natural Language Processing in Medicine: An Overview.* Methods of Information in Medicine 1996; 35: 285-301

[Stearns 01] Stearns MQ, Price C, Spackman KA, Wang AY (2001). *SNOMED Clinical terms: Overview of the Development Process and Project Status.* Proceedings of the 2001 AMIA Annual Fall Symposium, 662-666

[Stiller 02] Stiller H, Elsner-Ehrling U, Leititis JU (2002). *Maßnahmen zur Anpassung der Dokumentation für das künftige pauschalierende Entgeltsystem (AR-DRG-System) im Krankenhaus.* In: *DRG-Einführung in deutschen Krankenhäusern.* Sonderbeilage „Das Krankenhaus“, Kohlhammer

[Tsui 01] Tsui F-C, Wagner MM, Dato V, Chang C-CH (2001). *Value of ICD-9-Coded Chief Complaints for Detection of Epidemics.* Proceedings of the 2001 AMIA Annual Fall Symposium, 711-715

[Wang 01] Wang AY, Barret JW, Bentley T, Markwell D, Price C, Spackman KA, Stearns MQ (2001). *Mapping Between SNOMED RT and Clinical Terms Version 3: A Key Component of the SNOMED CT Development Process.* Proceedings of the 2001 AMIA Annual Fall Symposium, 741-745

[Wiesman 97] Wiesman F, Hasman A, van den Herik HJ (1997). *Information retrieval: an overview of system characteristics.* International Journal of Medical Informatics 47: 5-26

[Wingert 84] Wingert F (1984). *SNOMED – Systematisierte Nomenklatur der Medizin – SNOMED Manual.* Springer, Berlin

[Wren 02] Wren JD, Garner HR (2002). *Heuristics for Identification of Acronym-Definition Patterns Within Text: Towards an Automated Construction of Comprehensive Acronym-Definition Dictionaries.* Methods of Information in Medicine, zur Veröffentlichung angenommen

[Zaiß 96] Zaiß A, Schulz S, Graubner B, Klar R (1996). *Conversion Table between ICD-9 and ICD-10.* Medical Informatics Europe 1996, 193-197

[Zaiß 02] Zaiß A, Graubner B, Ingenerf J, Leiner F, Lochmann U, Schopen M, Schrader U, Schulz S (2002). *Medizinische Dokumentation, Terminologie und Linguistik.* In: Lehmann TM, Meyer zu Bexten E (Hg.), Handbuch der Medizinischen Informatik, Carl Hanser Verlag München Wien, 45-102

Danksagung

Herrn Prof. Rüdiger Klar und Herrn Prof. Udo Hahn danke ich für die Begutachtung dieser Arbeit. Darüber hinaus möchte ich mich bei Herrn Prof. Rüdiger Klar für die unermüdliche Unterstützung und Ermutigung bedanken, mit der er mich bei der Verfassung dieser Dissertation begleitet hat.

Für die Anregung zur Bearbeitung dieses Themas sowie für die geduldige Betreuung der Entwicklungsphase dieser Arbeit und das mir entgegengebrachte Vertrauen möchte ich mich eindringlich bei Herrn Dr. Albrecht Zaiß bedanken. Mein ganz besonderer Dank gilt Herrn Dr. Stefan Schulz, der meine Arbeit von den ersten Ansätzen bis zum Abschluss mit großem Interesse verfolgt und in intensiven Diskussionen befruchtet hat. Ohne ihn wäre die Publikation von Teilen der Arbeit nicht möglich gewesen. Seinem Engagement ist es auch zu verdanken, dass wertvolle Teile des Algorithmus mit anderen Projekten der Institute für Medizinischen Informatik und der Computerlinguistik verknüpft wurden und in ihnen weiterleben.

Für ihre großzügige Hilfsbereitschaft bei computertechnischen Problemen gilt mein Dank vor allem Frau Friedlinde Bühler, die für die nötigen Voraussetzungen zur praktischen Umsetzung meiner Arbeit sorgte. Darüber hinaus hat mich das freundliche Interesse vieler Mitarbeiter des Institutes für Medizinische Informatik immer wieder in meinen Anstrengungen bestärkt.

Herzlich danken möchte ich auch meinen Freunden in der WG für alle Einschränkungen, welche die Schreibphase mit sich brachte, und für die Geduld, mit der sie meine Arbeit mit ins Gebet genommen haben.

Ohne die menschliche Unterstützung und das Refugium, das mir meine Eltern Hedwig und Rudolf Franz immer wieder gewährten, wären weder mein Medizinstudium noch diese Promotion möglich gewesen. Für ihr Vertrauen und ihre Liebe gilt ihnen mein ganzer Dank.