Aus dem Fachbereich Medizin
der Johann Wolfgang Goethe-Universität
Frankfurt am Main
Institut für Dokumentation und Informationstechnologie

Automatische Diagnosekodierung mit XDIAG

Konzeption und Evaluation eines heuristischen Verfahrens zur leitbegrifforientierten automatischen Diagnosekodierung auf Basis der Daten des ICD10-Diagnosen-Thesaurus

Dissertation
zur Erlangung des Doktorgrades der theoretischen Medizin des Fachbereichs Medizin der Johann Wolfgang Goethe-Universität Frankfurt am Main

vorgelegt von Ralf Starzetz
aus Bad Camberg

Bad Camberg, 2004

Aus dem Fachbereich Medizin
der Johann Wolfgang Goethe-Universität
Frankfurt am Main
Institut für Dokumentation und Informationstechnologie

Automatische Diagnosekodierung mit XDIAG

Konzeption und Evaluation eines heuristischen Verfahrens zur leitbegrifforientierten automatischen Diagnosekodierung auf Basis der Daten des ICD10-Diagnosen-Thesaurus

Dissertation
zur Erlangung des Doktorgrades der theoretischen Medizin des Fachbereichs Medizin der Johann Wolfgang Goethe-Universität Frankfurt am Main

vorgelegt von Ralf Starzetz
aus Bad Camberg

Bad Camberg, 2004

Dekan: Prof. Dr. Josef Pfeilschifter

Referent: Prof. em. Dr. Wolfgang Giere

Koreferent: Prof. Dr. Andreas Goldschmidt

Tag der mündlichen Prüfung:

01. Juni 2005

Danksagung

Ich möchte mich an dieser Stelle ganz besonders herzlich bei Herrn Prof. em. Dr. med. Wolfgang Giere bedanken, der als mein Betreuer maßgeblich zum Gelingen der vorliegenden Arbeit beigetragen hat. Auch nach seinem Ausscheiden aus dem ZINFO Frankfurt hat er trotz zahlreicher anderer Verpflichtungen für Fragen und Probleme stets ein offenes Ohr gehabt und die Ausarbeitung meiner Ergebnisse mit Rat und Tat „bis zum letzten Pinselstrich“ begleitet.

Mein ganz besonderer Dank gilt weiterhin Frau Helga Schalck. Durch die Möglichkeit, zusammen mit Herrn Prof. Giere die wissenschaftlichen Ausarbeitungen ihres Mannes, Herrn Detlef Schalck, zu übernehmen, eröffnete sich mir der Rahmen, in dem die Ergebnisse der vorliegenden Dissertation erarbeitet werden konnten. Ich hoffe, daß meine Arbeit ein Stück dazu beiträgt, die besondere wissenschaftliche Leistung von Herrn Detlef Schalck zu würdigen. Ohne seine über Jahre gewachsenen konzeptionellen Vorarbeiten hätte die vorliegende Dissertation nicht geschrieben werden können.

Gerne denke ich auch an die Zeit am ZINFO Frankfurt zurück. Der in dieser Abteilung gepflegte Geist und die besondere wissenschaftliche Ausrichtung waren wichtige Grundlagen für das Gelingen der vorliegenden Arbeit.

Besonders erwähnen möchte ich hierbei Frau Marianne Wohlfahrt sowie Herrn Dr. rer. med. Wolfgang Kirsten: Frau Wohlfahrt hat mich mit den Gedanken von Herrn Schalck im Zusammenhang mit automatischer Diagnosekodierung in Kontakt gebracht hat und mir mit viel Geduld notwendiges Hintergrundwissen erläutert. Herr Dr. Kirsten hat mich bei der Erstellung der Arbeit kontinuierlich mit viel Einsatz und Kompetenz auch über rein inhaltliche Fragestellungen hinweg betreut. Beiden gilt mein herzlicher Dank.

Weiterhin möchte ich mich bei meinen Kollegen Andrea Volle, Dr. Christoph Winkler, Jörg Bay sowie Dr. Octavian Weiser bedanken, die durch ihre kompetente und kollegiale Unterstützung erheblich zu einer angenehmen und produktiven Arbeitsatmosphäre beigetragen haben. Besonderer Dank gebührt meinem Kollegen und Freund, Herrn Dipl.-Inf. Thorsten Diehl, der für mich die Umsetzung der Algorithmen des Prototypen übernommen hat. Ohne seine kompetente Hilfe hätte die vorliegende Arbeit nicht in dieser Form realisiert werden können.

Abschließend möchte ich mich auch ganz besonders herzlich bei meiner Familie sowie meiner langjährigen Freundin Tatjana Nattermann für deren Unterstützung, Geduld und Verständnis bedanken. Diese Arbeit widme ich meinen Vater.

I. Inhaltsverzeichnis

II. ABBILDUNGSVERZEICHNIS

III. TABELLENVERZEICHNIS

1. Einleitung

„Krankendaten: Dokumentation für Medizin oder Bürokratie?“

Dieser Titel steht am Anfang eines im Jahre 1984 von *Giere*[1] verfaßten Artikels, in dem in leicht provozierender Weise das Spannungsfeld aufgezeigt wird, in dem medizinische Dokumentation in der ärztlichen Praxis wahrgenommen wird: Medizinische Zweckmäßigkeit auf der einen und Bürokratie auf der anderen Seite bilden hierbei die Eckpfeiler eben dieses Spannungsfeldes.

Klassifikation und Kodierung – Themen, die wie kaum andere im medizinischen Alltagsleben ambivalent betrachtet werden und bei den unmittelbar davon Betroffenen durchaus gemischte Gefühle auslösen: Klassifikation kostet Zeit – Klassifikation bedeutet Mehrarbeit. Gleichwohl schafft Klassifikation aber auch die Basis für neue Erkenntnisse und hilft, bestehende Lehrmeinungen zu evaluieren.

In der vorliegenden Arbeit wird ein IT-basiertes[2] Verfahren zur automatischen Klassifikation von Diagnosen[3] vorgestellt, das eine Symbiose unterschiedlicher Ideen und Erfahrungen repräsentiert: Einerseits sind dies Ideen, die sich als Ergebnis jahrelanger intensiver und praxisnaher Beschäftigung mit medizinischer Dokumentation ergeben haben; andererseits werden Erfahrungen herangezogen, die im Verlaufe eines Projektes zur prototypischen Realisierung einer automatischen Diagnosekodierung gemacht wurden.
Das in den folgenden Abschnitten vorgestellte und diskutierte Verfahren ist durch seine klare und enge Zielsetzung ein ressourcensparendes Verfahren. Es stellt eine Kombination unterschiedlicher Basistechniken aus dem Bereich Informationsmanagement

---

1 Vgl. [GIERE 84].

2 Im Sinne einer besseren Lesbarkeit der vorliegenden Arbeit soll der Begriff „Informationstechnologie“, wie in der Praxis üblich, im folgenden mit „IT“ abgekürzt werden.

3 In der vorliegenden Arbeit wird das entwickelte Verfahren auf die Bearbeitung von Diagnosen eingeschränkt. Eine Erweiterung auf Prozeduren dürfte aber auf Grund der besonderen Eigenschaften des Verfahrens mit geringem Aufwand realisierbar sein.

dar. Durch den Verzicht auf die explizite Isolierung und Visualisierung einzelner Sachverhalte aus der Krankengeschichte erfolgt zwar einerseits eine völlige Abkehr von dem Konzept einer Meta-Krankengeschichte[4], anderseits kann auf diese Weise aber auf die Verwendung komplexer Regeln und somit auch auf die Verwendung komplexer Stammdaten verzichtet werden.

Das vorgestellte Verfahren ist nur und ausschließlich auf die Ermittlung von Diagnosekodes auf Basis einer Auswertung[5] vorliegender medizinischer Freitexte ausgerichtet. Als Stammdatenbasis werden Datenbestände verwendet, die entweder besonders leicht zu pflegen sind oder aber ohnehin permanent im Rahmen von Langzeitprojekten[6] gepflegt werden.

Auf diese Weise soll der Versuch demonstriert werden, aus medizinischen Texten durch geschickte, zielgerichtete und ressourcensparende Analyse ein Maximum an kodierrelevanter Information zu extrahieren und in Diagnosekodes umzusetzen.[7]

1.1 Ziele, Hintergründe und Hypothese

In den folgenden Abschnitten wird zunächst der Vorgang der Diagnosekodierung in einen größeren Bedeutungszusammenhang eingebettet. Insbesondere zwei Facetten sind hierbei von besonderer Wichtigkeit:

- Diagnosekodierung spielt eine zentrale Rolle im Rahmen medizinischer Dokumentation. Gerade medizinische Dokumentation ist aber bestimmt von einer zunehmenden Anzahl gesetzlich-organisatorischer Rahmenbedingungen.

- Diagnosekodierung hat darüber hinaus eine erhebliche Bedeutung für übergeordnete Prozesse zur medizinischen Wissensgenerierung.

---

4 Vgl. [GREGORI 95] sowie [LUZ 97].

5 Für die Zwecke der vorliegenden Arbeit soll diese Auswertung als „Kodierung“ bezeichnet werden. Im Sinne einer größeren Präzision müßte man an dieser Stelle eigentlich von „Meta-Kodierung“ sprechen (vgl. [FEIGL 74]), da in den zu bearbeitenden medizinischen Freitexten bereits Informationen „kodiert“ wurden.

6 Vgl. hierzu insbesondere die Ausführungen im Vorwort zu [DIMDI 01a].

7 Bei *Rector* [RECTOR 99] sowie bei *Cimino* [CIMINO 98] findet man in besonders eindrucksvoller Form einen Überblick über die komplexe Problematik im Rahmen der Erstellung einer „allgemeingültigen“ medizinischen Terminologie. *Spyns* [SPYNS 96] bietet einen Überblick über Methoden zur Verarbeitung natürlicher Sprache in der Medizin.
An dieser Stelle läßt sich somit leicht ableiten, warum der Fokus für die Ausführungen der vorliegenden Arbeit in der beschriebenen Art eingeschränkt wurde. *De Bruijn* [DEBRUIJN 98] grenzt darüber hinaus in besonders deutlicher Weise die Möglichkeiten und Perspektiven von Systemen zur Ermittlung von Kodes gegenüber Systemen zur vollständigen Sprachanalyse ab.

Eine Diskussion vorstehend beschriebener Punkte spannt den Bedeutungskontext auf, in dem sich das in der vorliegenden Arbeit diskutierte Verfahren positionieren läßt. Mit Hilfe dieses Bedeutungskontextes wird in einem nächsten Schritt die konkrete Zielsetzung der vorliegenden Arbeit aufgezeigt. Die hieraus abgeleitete Hypothese schließt das vorliegende Kapitel ab.

1.1.1 Diagnosekodierung als essentieller Bestandteil medizinischer Dokumentationspflicht

Strukturierte Aufzeichnungen haben in der Medizin eine lange Tradition. Seit *Hippokrates* kennt man die Dokumentation ärztlichen Wirkens in mehr oder weniger ausführlicher bzw. vollständiger Form. *Weed*[8] fordert in diesem Zusammenhang sogar, daß das vom Arzt im Rahmen der Patientenbehandlung zu erstellende Krankenblatt die Qualität eines wissenschaftlichen Manuskriptes haben sollte.

Seit Mitte der achtziger Jahre erweiterte sich der Blickwinkel, unter dem medizinische Dokumentation gesehen wird, insbesondere durch folgende einschneidende Veränderungen der juristisch-organisatorischen Rahmenbedingungen des ärztlichen Wirkens und Handelns[9]:

- Änderung der ärztlichen Berufsordnung
- Änderung der medizinisch relevanten Rechtsprechung
- Änderung der Bundespflegesatzverordnung

Die von *Giere* aufgezeigten Änderungen führten in erster Konsequenz dazu, daß die Relevanz medizinischer Basisdokumentation an sich stieg.

Besonders deutlich läßt sich dies anhand eines Auszuges aus der ärztlichen Berufsordnung für Hessen aufzeigen.

---

8 Vgl. [WEED 78].
9 Vgl. [GIERE 86b].

In Paragraph 11, Satz 1 ist folgendes festgehalten[10]:

> **„Der Arzt hat über die in Ausübung seines Berufes gemachten Feststellungen und getroffenen Maßnahmen die erforderlichen Aufzeichnungen zu machen. Ärztliche Aufzeichnungen sind nicht nur Gedächtnisstützen für den Arzt, sie dienen auch dem Interesse des Patienten an einer ordnungsgemäßen Dokumentation."**

Vorstehend aufgeführte Formulierungen lassen, bedingt durch ihren grundlegenden Charakter, Klassifikation bzw. Kodierung als speziellen Bestandteil medizinischer Dokumentation zunächst unerwähnt.

Folgende ganz aktuelle organisatorische Rahmenbedingungen des praktischen ärztlichen Handelns führen über den grundlegenden Begriff der „Dokumentation" hinaus zur konkreten Forderung nach einer systematischen „Kodierung" medizinischer Diagnosen[11]:

Abrechnung der im Krankenhaus erbrachten Leistungen nach Fallpauschalen und Sonderentgelten

Die Basis der Abrechnung nach Fallpauschalen und Sonderentgelten ist die bei der Entlassung des Patienten zu kodierende Hauptdiagnose[12].

Abrechnung der im Krankenhaus erbrachten Leistungen nach DRGs[13]

Mit der Einführung der Abrechnung auf Basis der DRGs gewinnt die vollständige und korrekte Kodierung medizinischer Diagnosen eine gegenüber dem System der Fallpauschalen und Sonderentgelte nochmals gestärkte Bedeutung:

Neben der zu kodierenden Hauptdiagnose spielen bei der Abrechnung auf Basis der DRGs gerade auch die kodierten Nebendiagnosen eine wichtige Rolle, da diese beispielsweise den Schweregrad einer Erkrankung determinieren und somit in der Regel den abzurechnenden Aufwand mitbestimmen.

---

[10] Vgl. [BO 85].
[11] Vgl. [ZAISS 02].
[12] Gemäß WHO (World Health Organization) ist die Hauptdiagnose diejenige Diagnose, die während der Behandlung des Patienten die meisten Ressourcen verbraucht hat.
[13] DRG = Diagnosis Related Groups; einen Überblick über das neue Abrechnungssystem findet man beispielsweise bei *Lauterbach* [LAUTERBACH 00].

Man erkennt, daß gerade die beiden vorstehend beschriebenen aktuellen Faktoren insbesondere auch einen betriebswirtschaftlich orientierten Anreiz zu vollständiger und korrekter Diagnosekodierung bieten.

1.1.2 Diagnosekodierung und medizinischer Erkenntnisgewinn im BAIK-Modell

Nachdem im vorstehenden Abschnitt insbesondere juristische, organisatorische und wirtschaftliche Bestimmungsfaktoren der medizinischer Diagnosekodierung aufgezeigt wurden, soll nunmehr anhand des BAIK-Modells[14] demonstriert werden, daß darüber hinaus die Kodierung medizinischer Diagnosen und somit deren Klassifikation einen essentiellen Schritt im Rahmen des medizinischen Erkenntnisfortschritts darstellt.

Kodierung muß somit als ein zentraler Unterstützungsfaktor für eine qualifizierte Patientenversorgung sowie für medizinische Forschung und Lehre betrachtet werden.

Um diesen Unterstützungsfaktor angemessen motivieren zu können, erscheint es zunächst sinnvoll, einen kurze Beschreibung des BAIK-Modells wiederzugeben:

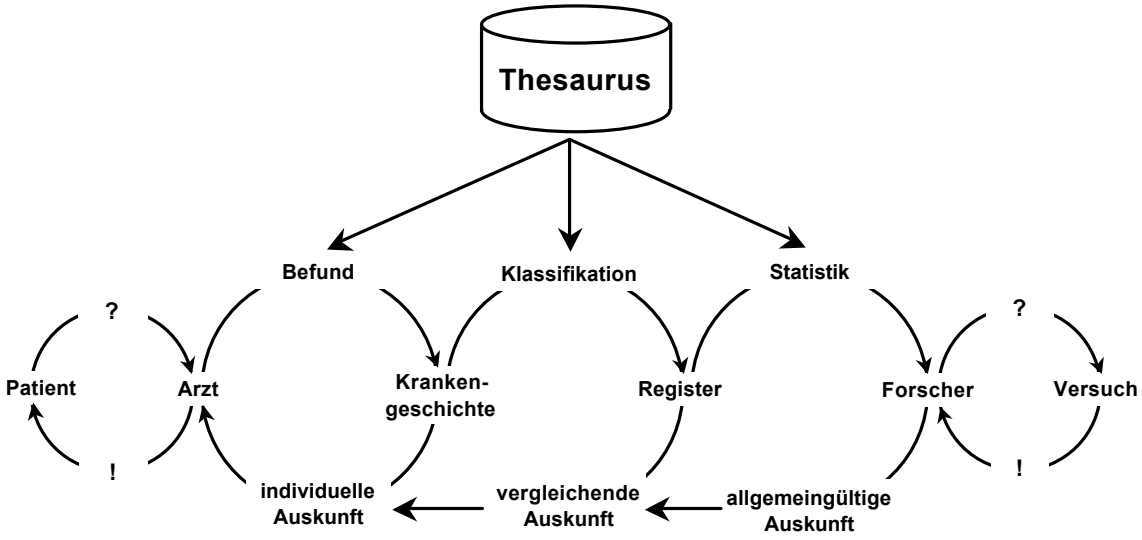

Abbildung 1: Das BAIK-Modell

[14] Vgl. [GIERE 86a].

Das von *Giere* entwickelte BAIK-Modell modelliert den Informationsfluß, der sich im Rahmen der Behandlung von Patienten durch Ärzte ergibt. Hierbei lassen sich drei Informationszyklen unterscheiden:

- Der Behandlungsorientierte Informationszyklus
- Der Vergleichsorientierte Informationszyklus
- Der Erkenntnisorientierte Informationszyklus

Der Behandlungsorientierte Informationszyklus

Ein Patient kommt mit einem Problem zum Arzt. Der Arzt erhebt Befunde und trägt diese in die Krankenakte des Patienten ein. Die in diesem Schritt entstandene Dokumentation ist patientengebunden. Die Auskunft bzw. Information, die hieraus zum Zwecke der Behandlung des Patienten gewonnen werden kann, ist **individuell**.

Der Vergleichsorientierte Informationszyklus

Die in der patientenindividuellen Dokumentation gesammelten Befunde eines Patienten lassen sich auf Basis eines vorgegebenen Ordnungsschemas (Klassifikationsschemas) klassifizieren. Es läßt sich auf diese Weise eine standardisierte Dokumentation erstellen. Diese Standardisierung erlaubt eine mit anderen ebenfalls standardisierten Fällen **vergleichende** Auskunft. Dieser anhand der Klassifikation mögliche Vergleichsprozeß erlaubt es dem Arzt, Zusatzinformationen zum aktuellen Patienten zu gewinnen.

Der Erkenntnisorientierte Informationszyklus

Auf Basis der Daten der standardisierten Dokumentationen lassen sich durch statistische Auswertung durch die Klassifikation vergleichbar gewordener Fälle Informationen gewinnen, die es einem Forscher ermöglichen, Hypothesen zu formulieren und diese anschließend zu bestätigen oder zu widerlegen. Die Abläufe an dieser Stelle können im Sinne einer **allgemeingültigen** Auskunft verstanden werden.

Man erkennt leicht, daß der Prozeß der Klassifikation und somit der Prozeß der Diagnosekodierung eine wichtige Rolle im Rahmen des vorgestellten Modells spielt. Nur durch Klassifikation werden individuell erfaßte Fälle vergleichbar und somit auf breiter Basis effizient und effektiv auswertbar.[15]

Es sei an dieser Stelle deutlich festgehalten, daß das im Rahmen der vorliegenden Arbeit vorgestellte Verfahren im Informationsfluß des BAIK-Modells in den Vergleichsorientierten Zyklus einzubetten wäre.

Man erkennt leicht, warum gerade das BAIK-Modell sich in hervorragender Weise eignet, die Ausführungen der vorliegenden Arbeit in einen größeren Bedeutungszusammenhang zu stellen:
Durch die explizite Darstellung der Informationsflüsse wird die zentrale Rolle der Kodierung medizinischer Diagnosen deutlich hervorgehoben.

In der Literatur findet man zwar auch andere Darstellungen, die die Notwendigkeit der „Wiederbenutzbarkeit“ von Daten und somit die Verbindung zwischen klinischer Praxis, medizinischer Forschung und Lehre aufzeigen, jedoch wird die besondere Rolle der Diagnosekodierung nicht so transparent wie im Rahmen des BAIK-Modells herausgearbeitet.

1.1.3 Diagnosekodierung in der medizinischen Praxis

Traditionell wird in Deutschland die Diagnosekodierung von medizinisch versiertem Fachpersonal *neben* der eigentlichen Tätigkeit durchgeführt. Hierzu kommen beispielsweise Ärzte oder in der medizinischen Dokumentation besonders geschulte verwaltungsorientierte Assistenten in Frage. Gerade diese in der Praxis häufig anzutreffende Organisationsform ist aber mit erheblichen Problemen behaftet:
Durch die erhebliche Arbeitsbelastung der Betroffenen bleibt für Kodieraufgaben häufig nur wenig Zeit.[16] Hieraus können Qualitätsprobleme bezüglich der ermittelten Diagnosekodes resultieren.[17] Weiterhin ergeben sich Probleme beim interpersonellen Vergleich[18] der Ergebnisse sowie bei deren Reproduzierbarkeit.

---

[15] *Ertel* [ERTEL 73] weist in diesem Zusammenhang insbesondere auf die Möglichkeit hin, bei entsprechender Aufbereitung und Repräsentation medizinische Informationen *automatisch* verknüpfen zu können.

[16] Vgl. [FEIGL 73].

[17] Vgl. zu Kriterien der Dokumentationsqualität die Ausführungen in [ZAISS 02]. Vgl. zu typischen Fehlern die Ausführungen in [LLOYD 85] sowie in [NITZSCHKE 92].

[18] Vgl. [DEBRUIJN 98].

Hall [19] zählt darüber hinaus zusammenfassend folgende idealtypische Gruppen von Fehlern im Rahmen der manuellen Diagnosekodierung auf:

- **Faktisch korrekte aber nutzlose Kodes**
  Beispiel: Alle gutartigen Läsionen werden als „kein Tumor“ kodiert.

- **Inkonsistente Kodierung**
  Beispiel: „Dysplasie“ an einem Tag / „Atypie“ an einem anderen Tag.

- **Ungewöhnliche oder nicht nachvollziehbare Abkürzungen**

- **Reine Eingabefehler**
  Beispiel: Eingabe von „Lipom“ an Stelle von „Lymphom“.

Alle vorgenannten Punkte stellen die Vollständigkeit und Korrektheit der in der Praxis manuell erzielten Kodierergebnisse und somit deren Wert für den medizinischen Erkenntnisfortschritt permanent in Frage.[20]

Die vorstehend kurz umrissenen Probleme zeigen deutlich den Bedarf nach geeigneten Instrumenten zur Unterstützung bei der Bewältigung praktischer medizinischer Kodieraufgaben auf.
In diesem Zusammenhang lassen sich zunächst grob zwei wichtige Unterstützungsformen abgrenzen:

Gedruckte Medien

Traditionell spielen in der Medizin gedruckte Medien eine große Rolle. Man denke in diesem Zusammenhang nur an Lehrbücher oder Tabellenwerke. Für medizinische Kodieraufgaben lassen sich aktuell als wichtige gedruckte Hilfsmittel folgende Bücher anführen:

- Die drei Bände der ICD-10 stellen das Basisgerüst der verwendeten Kodiersystematik dar.[21]

---

[19] Vgl. [HALL 86].
[20] Einen ausführlichen vergleichenden Überblick über medizinische Kodiersysteme und deren Anspruch bzw. Zielsetzung findet man bei *Cimino* [CIMINO 96].
[21] Vgl. [ICD10 03a], [ICD10 03b] sowie [ICD10 99].

- Der ICD-10-Diagnosenthesaurus (IDT)[22] schlägt die Brücke zum täglichen praktischen Sprachgebrauch und ist somit bei der praktischen Diagnosekodierung nicht mehr wegzudenken.

Der Einsatz von Informationstechnologie im Rahmen medizinischer Kodieraufgaben

Giere[23] weist bereits 1969 darauf hin, daß eine besondere Chance für den sinnvollen Einsatz von Informationstechnologie im Krankenhaus in der Vereinfachung von Verwaltungsroutinen besteht.[24] Diese Einschätzung bleibt auch oder gerade bei Berücksichtigung der Entwicklung der IT im medizinischen Umfeld aktuell:

Giere charakterisiert im Jahre 2002[25] die automatische Textanalyse und Klassifikation als wichtigen Prüfstein bei der Beurteilung von IT-Systemen zur Realisierung digitaler Patientenakten.

Es liegt somit grundsätzlich auf der Hand, daß Diagnosekodierung eine Verwaltungsroutine ist, deren Bewältigung durch den Einsatz von Computern sinnvoll zu unterstützen ist.[26]

Das Spektrum der in diesem Zusammenhang möglichen Lösungsansätze ist breit gefächert. Als Randpunkte eben dieses Spektrums lassen sich nennen:

- IT als Kodier*unterstützung*
- IT-gestützte *automatische* Kodierung

Das auf diese Weise abgegrenzte Spektrum soll im folgenden näher beschrieben werden:

---

22 Vgl. [DIMDI 01].
23 Vgl. [GIERE 69].
24 Vgl. hierzu insbesondere auch die Ausführungen von *Talmon* [TALMON 02], der in sehr anschaulicher Weise aktuelle Probleme und Chancen der Medizinischen Informatik diskutiert sowie die grundlegenden Aussagen von *Sittig* [SITTIG 94].
25 Vgl. [GIERE 02].
26 Eine eindrucksvolle Demonstration des Nutzens IT-orientierter Hilfsmittel im Rahmen medizinischer Diagnosekodierung findet man bei *Hohnloser et al.* [HOHNLOSER 96] sowie [HOHNLOSER 96a].

Varianten der Kodierunterstützung durch IT

IT als modernes Präsentationsmedium

In zahlreichen medizinisch orientierten IT-Anwendungen werden ursprünglich in gedruckter Form vorliegende Informationen zusätzlich in Form von Text- oder Grafikdateien auf Computern zur Verfügung gestellt. Es ändert sich somit zunächst nur das zur Präsentation der Unterstützungsinformationen herangezogene Medium. Werden die im Computer vorliegenden Daten um qualifizierte Such- und Filterfunktionen erweitert, so ergibt sich eine nächste Kategorie.

IT als flexibles, individuelles und situationsgerechtes Präsentationsmedium

Durch die Implementierung qualifizierter interaktiver Such- und Filterfunktionen lassen sich IT-Systeme sehr sinnvoll im Rahmen medizinischer Kodieraufgaben einsetzen. Ein besonders gutes Beispiel für diese Art von Hilfsmitteln ist die Realisierung von sogenannten „Kodierungsbrowsern“[27]. Derartige Browser können dem Benutzer helfen, die benötigten Informationen schrittweise interaktiv zu erarbeiten und hierbei stets den medizinisch relevanten Kontext im Auge zu behalten.

In diesem Zusammenhang sei auch besonders auf die Möglichkeit des Einsatzes von „Hyperlinks“[28] hingewiesen, mit deren Hilfe eine umfangreiche Informationsmenge sinnvoll strukturiert werden kann. Durch Querverweise erhält der Benutzer beispielsweise Hinweise auf aktuelle kodierrelevante Zusatzinformationen oder auf geeignete Alternativkodes.[29]

Die vorstehend beschriebenen Ansätze sind allesamt dadurch gekennzeichnet, daß letztlich der Benutzer, vom Computer *unterstützt*, die Entscheidung über den jeweils auszuwählenden Kode trifft. Entfällt diese letzte Entscheidung, so ergibt sich die abschließende Kategorie.

---

[27] Der insbesondere aus dem Bereich des Internet bekannte Begriff des „Browsers“ kennzeichnet im vorstehenden Zusammenhang die IT-gestützte Realisierung von Betrachtungs- und Suchfunktionen.

[28] Für die Zwecke der vorliegenden Arbeit soll der Begriff „Hyperlink“ im Sinne einer „intelligenten Verknüpfung“ verstanden werden.

[29] Eine Diskussion eines bestehenden Systems zur IT-basierten klinischen Kodierunterstützung findet man beispielsweise bei *Bouchet* [BOUCHET 98] [BOUCHET 98a].
Weitere in diesem Zusammenhang relevante Erfahrungsberichte findet man beispielsweise bei *Deimel* [DEIMEL 97], *Lloyd* [LLOYD 97] sowie *Kuchenbecker* [KUCHENBECKER 95].

IT als Träger eines automatischen Kodierverfahrens

Hält man sich die Ausführungen der vorstehenden Abschnitte vor Augen, so läßt sich für den Begriff der automatischen Diagnosekodierung eine Definition ableiten, die dieses Konzept im Rahmen der vorliegenden Arbeit charakterisieren soll:[30]

> **„Automatische Diagnosekodierung soll genau dann vorliegen, wenn der Kodierungsprozeß durch die Implementierung eines oder mehrerer Algorithmen auf einem IT-System realisiert und somit der Vorgang der Kodierung an sich für den Benutzer transparent ist. Der Computer kodiert die bereitgestellten Informationen und fordert vom Benutzer ggf. ergänzende oder fehlende Angaben an.**
> **Weiterhin soll davon ausgegangen werden, daß die zur automatischen Diagnosekodierung erforderlichen Informationen in einer für den oder die Algorithmen geeigneten Form vorliegen[31], so daß sich die Eingriffe des Benutzers auf die Bereitstellung dieser Eingangsinformationen sowie auf die Weiterverarbeitung der erstellten Kodes beschränken."**

1.1.4 Automatische Diagnosekodierung als heuristischer Ansatz

Nachdem im vorstehenden Abschnitt die IT-gestützte automatische Diagnosekodierung diskutiert und für die Zwecke der vorliegenden Arbeit definiert wurde, soll nunmehr der Ansatz präzisiert werden, um auf diese Weise die Grundlage für die nachfolgende Hypothese zu legen.

Hält man sich die Komplexität der deutschen Sprache vor Augen und berücksichtigt man weiterhin, daß die medizinische Fachsprache eine Teilmenge der deutschen Sprache repräsentiert, so läßt sich leicht folgern, daß eine Bearbeitung medizinischer Texte mit dem Ziel, deren Inhalt bzw. Inhalte zu klassifizieren, eine komplexe Herausforderung darstellt. Weiterhin liegt somit auf der Hand, daß für eine perfekte medizinische Diagnosekodierung grundsätzlich neben einem reichen Erfahrungsschatz ein vollständiges Verständnis der zugrunde liegenden Informationen unabdingbar ist.

---

[30] Vgl. [MOORE 94].
[31] Zur besonderen Eignung medizinischer Routinedokumente siehe insbesondere [RÖTTGER 73].

Es läßt sich folgern, daß für eine fehlerfreie *automatische* Diagnosekodierung theoretisch ein vollständiges Verständnis der entsprechenden Informationen durch den Computer erfolgen muß. Für die automatische Diagnosekodierung auf Basis medizinischer Texte bedeutet dies in der Praxis, daß im Rahmen der Implementierung entsprechender Algorithmen ein vollständiges Textverständnis mit Erschließung der syntaktischen, semantischen und pragmatischen Ebenen realisiert werden muß.

Die Praxis der Klartextanalyse in der Medizin hat gezeigt[32], daß selbst mit modernsten linguistischen Verfahren unter Einsatz modernster Technik ein derartiges Textverständnis zur Zeit als problematisch zu betrachten ist. Eine Klartextverarbeitung auf einem gewissen Erkennungsniveau ist zwar möglich, doch nichtverarbeitbare „Grauzonen“ bleiben und sind Gegenstand der aktuellen medizinlinguistischen Forschung.

Im Rahmen der vorliegenden Arbeit wird an Stelle komplexer Textanalyse ein Ansatz verfolgt, der Schnelligkeit, Wartbarkeit und unbedingte Praxisorientierung in den Vordergrund stellt. Ganz ausdrücklich ist es, wie bereits beschrieben, hierbei nicht das Ziel des vorgestellten Konzepts, medizinische Diagnosetexte zu verstehen bzw. inhaltlich vollständig zu erschließen. Vielmehr soll der Versuch demonstriert werden, aus medizinischen Texten durch geschickte Analyse ein Maximum an kodierrelevanter Information zu extrahieren und aufzuarbeiten. Auf diese Weise ist stets sichergestellt, daß der Benutzer gemäß den Anforderungen *Van Bemmels*[33] durch den Einsatz von IT eine bessere Entscheidungssituation vorfindet als ohne den Einsatz von IT.

Die Umsetzung der vorstehend beschriebenen Zielsetzung läßt sich somit als Versuch der Realisierung einer **Heuristik** charakterisieren. Der Begriff der Heuristik wird in der Regel im Bereich des Operations-Research[34] verwendet und muß aus diesem Grunde für die Zwecke der vorliegenden Arbeit zunächst geeignet definiert werden.

In Erweiterung eben dieser Definition soll anschließend ein exakter begrifflicher Rahmen für die Hypothese sowie die spätere kritische Würdigung der vorgestellten Lösungsansätze bereitgestellt werden.

---

[32] Vgl. [ZAISS 02].
[33] Vgl. [VANBEMMEL 00].
[34] Teilgebiet der Betriebswirtschaftslehre, das sich mit mathematischer Planungsrechnung insbesondere in den Bereichen Produktionsplanung und Ablauforganisation beschäftigt.

Als Basis der nachfolgenden Ausführungen wird der Begriff der „Heuristik“ somit folgendermaßen zweckorientiert definiert:

> **„Heuristische Verfahren bieten keine Garantie, daß eine optimale Lösung gefunden wird. Sie beinhalten lediglich bestimmte Vorgehensregeln, die für die jeweilige Problemstruktur sinnvoll und erfolgversprechend sind.“**[35]

Man unterscheidet insbesondere zwei Gruppen heuristischer Verfahren, die im Rahmen der vorliegenden Arbeit besonders relevant sind:

- **Eröffnungsverfahren**
- **Verbesserungsverfahren.**

Eröffnungsverfahren
Eröffnungsverfahren dienen zur Bestimmung einer ersten zulässigen Lösung.

Verbesserungsverfahren
Verbesserungsverfahren dienen zur Verbesserung einer ermittelten zulässigen Lösung. Bereits an dieser Stelle sei darauf hingewiesen, daß sich die Charakteristika eben dieser beiden Verfahren in der Konzeption des realisierten Prototypen widerspiegeln:
Man versucht zunächst, einen rudimentären aber grundsätzlich richtigen Kodierungsansatz zu finden und arbeitet anschließend an dessen Verbesserung bzw. Verfeinerung.

[35] Vgl. [DOMSCHKE 93].

1.1.5 Hypothese: Ein heuristisches Verfahren zur leitbegrifforientierten automatischen Diagnosekodierung auf Basis der Daten des ICD-10-Diagnosenthesaurus ist möglich und sinnvoll

Aus dem in den vorstehenden Abschnitten beschriebenen Bedeutungskontext sowie aus der hieraus abgeleiteten Zielsetzung für eine heuristische automatische Diagnosekodierung läßt sich folgende Arbeitshypothese ableiten, die als Leitlinie zur Evaluation des zu untersuchenden Verfahrens dienen soll:

> **Ein heuristisches leitbegrifforientiertes Verfahren zur IT-unterstützten automatischen Kodierung medizinischer Diagnosen, das sich am ICD-10-Diagnosen-Thesaurus als Stammdatenbasis orientiert, läßt sich effizient und effektiv in bestehende Dokumentationsabläufe einbetten. Die hierbei zu extrahierenden Informationen tragen gemäß BAIK-Modell zur Konsolidierung und Generierung medizinischen Wissens bei, so daß im Rahmen gesetzlicher vorgeschriebener Dokumentationsvorgänge gleichsam die Grundlage für substantiierten medizinischen Erkenntnisfortschritt gelegt werden kann.**

1.2 Aufbau der Arbeit

Ein besonders wichtiger Begriff im Rahmen der vorliegenden Arbeit ist der Begriff des „Thesaurus“. Wegen dieser grundlegenden Bedeutung für die nachfolgenden Ausführungen soll zunächst eine informationswissenschaftlich orientierte Begriffsdefinition nach DIN 1463 erfolgen:[36]

[36] Vgl. [BURKART 90].

„Ein Thesaurus ... ist eine geordnete Zusammenstellung von Begriffen und ihren Bezeichnungen, die in einem Dokumentationsgebiet zum Indexieren, Speichern und Wiederauffinden dient. Er ist durch folgende Merkmale gekennzeichnet:

I. Begriffe und Bezeichnungen werden eindeutig aufeinander bezogen („terminologische Kontrolle"), indem

I.a Synonyme möglichst vollständig erfaßt werden, ...

I.b für jeden Begriff eine Bezeichnung (z.B. Vorzugsbenennung[37]) festgelegt wird, die den Begriff eindeutig vertritt.

II. Beziehungen zwischen den Begriffen (repräsentiert durch ihre Bezeichnungen) werden dargestellt."

Der Aufbau der vorliegenden Arbeit orientiert sich an den praktisch notwendigen Schritten, um von einer Idee zu einer evaluierbaren Lösung mit dem Ziel der automatischen Diagnosekodierung zu kommen:
Am Anfang der Arbeit steht eine Einführung in die Thematik mit den für die weiteren Darstellungen notwendigen Begriffsbestimmungen. Anhand dieser Einführung läßt sich gleichsam die Motivation aufzeigen, aus der heraus die Beschäftigung mit Methoden zur automatischen Diagnosekodierung überhaupt erst erfolgt. Der einleitende Abschnitt mündet in die der Arbeit zugrundeliegende Hypothese.
Im hierauf folgenden Abschnitt werden die für den Lösungsansatz relevanten Determinanten herausgearbeitet: Die kodierrelevanten Aspekte medizinischer Texte stecken die sprachliche Domäne ab, innerhalb derer sich die nachfolgend dargestellten Ansätze und Ideen zu bewähren haben.
Nachdem nunmehr die Thematik grundsätzlich motiviert und das Arbeitsfeld auch begrifflich abgesteckt ist, erfolgt eine Einführung in die konzeptionellen Grundlagen der eingesetzten Lösungsverfahren. Diese Lösungsverfahren basieren auf Konzepten und Ideen, die dem Nachlaß von Herrn *Detlef Schalck* entnommen wurden. Um den an dieser Stelle für die Zwecke der vorliegenden Arbeit vorhandenen „Schatz" sinnvoll „heben" zu können, ist es zunächst notwendig, die Ideen *Schalcks* möglichst unverfälscht wiederzugeben. Diese Wiedergabe erfolgt im Rahmen einer idealtypischen Darstellung vorgeschlagener möglicher Verfahrensabläufe.

[37] Im Rahmen der vorliegenden Arbeit wird der Begriff „Vorzugsbenennung" synonym zu dem Begriff „Vorzugsbegriff" verwendet.

Ziel dieses Abschnittes ist es somit, das Konzept von *Schalck* möglichst transparent und in seinen Grundideen nachvollziehbar darzustellen, ohne an dieser Stelle die Darstellung durch Implementierungsaspekte unangemessen zu verwässern.
Gerade diese Implementierungsaspekte werden im nachfolgenden Abschnitt der vorliegenden Arbeit dargestellt. Es erfolgt an dieser Stelle also der Schritt von den Lösungsansätzen *Schalcks* hin zu konkreten Lösungsverfahren. Es liegt auf der Hand, daß in diesem Zusammenhang natürlich auch die hierbei benötigten Daten dargestellt werden müssen.

Die Darstellung der Lösungsalgorithmen abstrahiert vollständig von einer zu deren Realisierung verwendeten Programmiersprache und kann somit als universelle Grundlage einer praktischen Umsetzung verstanden werden.

Natürlich wurden die im Rahmen der Erstellung der vorliegenden Arbeit entwickelten Lösungsansätze mit Hilfe programmtechnischer Werkzeuge realisiert, um die Verfahren einer gründlichen Evaluation zuführen zu können.

Kern der abschließenden Evaluation ist daher ein gründlicher Abgleich der entwickelten Lösungsansätze mit den sich aus der Hypothese ergebenden Anforderungen. Hierbei wird insbesondere überprüft, ob die erarbeiteten Lösungsansätze den einführend aufgezeigten Charakteristika medizinischer Fachsprache gerecht werden.

Auf Basis eben dieser Ergebnisse kann ein Ausblick gewagt werden:

Im Rahmen einer abschließenden Diskussion der erarbeiteten und dargestellten Ergebnisse erfolgt eine zusammenfassende Würdigung des Konzeptes der Leitbegrifforientierung sowie der hieraus im Rahmen einer praktischen Realisierung erarbeiteten Verfahren (XDIAG)[38].

Eine Zusammenfassung sowie ein Ausblick auf mögliche Ergänzungen oder Weiterentwicklungen bilden den Abschluß der vorliegenden Arbeit.

[38] Aus Gründen der besseren Lesbarkeit soll in den folgenden Abschnitten der vorliegenden Arbeit mit dem Begriff „XDIAG“ stets das konzeptionell entwickelte und prototypisch realisierte Gesamtsystem zur automatischen leitbegrifforientierten Diagnosekodierung mit allen notwendigen Systembestandteilen (Daten und Algorithmen) bezeichnet werden.

2. Ausgewählte kodierrelevante Charakteristika medizinischer Texte

Im folgenden Abschnitt sollen nunmehr einige Charakteristika herausgearbeitet werden, die medizinische Texte in der Praxis kennzeichnen und somit auch für Verfahren im Rahmen der automatischen Diagnosekodierung relevant sind. Die Bewältigung der sich aus eben diesen Charakteristika ergebenden Herausforderungen im Rahmen konkreter Ansätze zur automatischen Diagnosekodierung können aus diesem Grunde als Determinanten der Entwicklung derartiger Ansätze herangezogen werden.[39]
Eben diese Charakteristika werden vor der Charakterisierung der konzeptionellen Grundlagen der Lösungsansätze sowie vor der Beschreibung des entwickelten Prototypen dargestellt, da nur auf diese Weise die konkret erarbeiteten Lösungsansätze praxisgerecht eingeführt, erläutert und diskutiert werden können.
Auch die Evaluation im Rahmen der vorliegenden Arbeit profitiert von den Ausführungen der folgenden Abschnitte:

Die Bewertung der Bewältigung der vorstehend beschriebenen Herausforderungen durch XDIAG stellt eine wesentliche Komponente des zum Zwecke der Evaluation aufgespannten Untersuchungsrahmens dar.

In den folgenden Abschnitten sollen nunmehr in Anlehnung an die Ausführungen von *Zaiss*[40] und *Lovis*[41] die Charakteristika der medizinischen Fachsprache herausgearbeitet werden. Hierbei wird zwischen formalen und inhaltlichen Charakteristika unterschieden. Formale Charakteristika beziehen sich eher auf Fragen im Zusammenhang mit Wortformen bzw. mit Wortbildung. Inhaltliche Charakteristika zeigen eher Probleme im Zusammenhang mit Wortbedeutungen auf.
Es muß an dieser Stelle festgehalten werden, daß die beiden Bereiche keinesfalls überschneidungsfrei sind. Die folgende Unterscheidung dient nur einer übersichtlicheren Darstellung. Nur bei simultaner Berücksichtigung formaler und inhaltlicher Aspekte kann medizinische Fachsprache in ihrer vollen Komplexität angemessen erfaßt werden.

---

[39] Vgl. hierzu insbesondere die Übersichtsartikel von *Wingert* [WINGERT 74] sowie *Schalck et al.* [SCHALCK 74].
[40] Vgl. [ZAISS 02].
[41] Vgl. [LOVIS 00].

2.1 Formale Charakteristika

Weitverbreitete Benutzung lateinischer und griechischer Wörter und Wortstämme

Medizinische Fachsprache ist geprägt von der intensiven Benutzung lateinischer und griechischer Wörter. Neben der Verwendung des „reinen“ Fremdsprachenvokabulars findet man auch zahlreiche hybride Wortbildungen, in denen beispielsweise deutsche und lateinische oder deutsche und griechische Wortstämme verschmelzen. Als Beispiel läßt sich „Ulnarislähmung“ anführen. In den letzen Jahren hat sich darüber hinaus eine Tendenz zur Verwendung englischer Wörter entwickelt, so daß die vorstehend für die lateinische und griechische Sprache erwähnten Phänomene zunehmend auch für die englische Sprache gelten.

Besonders hervorzuheben ist die Bedeutung von Präfixen oder Suffixen: Durch die Existenz zahlreicher Hybridwörter sowie die spezifische Bedeutung bestimmter Suffixe (z. B. „itis“ für Entzündungen) existieren in der medizinischen Fachsprache zahlreiche Wörter mit ähnlichen Präfixen oder Suffixen.

Vorstehend beschriebene Punkte zeigen deutlich, daß bei der medizinischen Textanalyse die orthographischen Charakteristika mehrerer Sprachen berücksichtigt werden müssen. Weiterhin müssen die verwendeten Algorithmen die hohe Zahl identischer Präfixe und Suffixe bewältigen.

Vorherrschende Sprachökonomie

In der medizinischen Fachsprache lässt sich überwiegend ein kompakter Telegrammstil beobachten. Folgende Merkmale lassen sich als charakteristisch für eben diesen Telegrammstil anführen:

- Tendenz zur Nominalisierung[42]
- Häufige Bildung und Verwendung von Komposita
- Häufiges Auftreten von Abkürzungen

[42] Der nominalsprachliche Charakter medizinischer Fachsprache ist ein eher inhaltliche ausgerichtetes Kriterium und wird aus diesem Grunde im Zusammenhang mit den inhaltlichen Charakteristika medizinischer Sprache diskutiert.

Häufige Bildung und Verwendung von Komposita

Bereits aus den vorstehenden Ausführungen im Zusammenhang mit Präfixen und Suffixen lässt sich ableiten, daß es in der Medizin eine ausgeprägte Tendenz zur Bildung von Komposita gibt. Das für die deutsche Sprache grundsätzliche Phänomen[43] der Wortkomposition tritt somit im medizinischen Umfeld noch verstärkt auf.
Eine analytische Erschließung eines Kompositums im Sinne eines Zugangs zu den in eben diesem Kompositum enthaltenen Einzelworten wird durch folgende in der medizinischen Fachsprache besonders ausgeprägte Phänomene erschwert:

- Vokalschwund (Beispiel: Nephr-ektomie)
- Konsonantenschwund (Beispiel: Sy-stole)
- Konsonantenassimilation (Beispiel: Ap-pendix)
- Einfügung von Vokalen (Beispiel: Hepat-o-splen-o-megalie)
- Einfügung von Konsonanten (Beispiel: A-n-ämie)

Man erkennt leicht, daß in den vorstehend aufgeführten Fällen beispielsweise ein einfacher Mustervergleich nicht zu einer korrekten Zerlegung der Komposita in Einzelworte führen kann.

Häufiges Auftreten von Abkürzungen

In medizinischen Texten finden sich zahlreichen Abkürzungen bzw. Akronyme. Auch dieses Phänomen ist sicher ein Resultat der vorherrschenden Spachökonomie. Eine konsequent einheitliche Verwendung von Abkürzungen würde für die analytische Erschließung medizinischer Texte sicher kein echtes Problem darstellen. Gleichwohl ist an dieser Stelle zu beachten, daß gerade bei Abkürzungen die Bandbreite orthographischer Variationen sehr groß ist, was den analytischen Zugang in der Praxis erheblich erschwert.
Besondere analytische Schwierigkeiten ergeben sich, wenn Abkürzungen nicht eindeutig sind. Da es sich in einem solchen Falle aber um eine inhaltliche Fragestellung handelt, wird der sich hieraus ergebende Fragenkomplex im Abschnitt „inhaltliche

[43] Man beachte in diesem Zusammenhang insbesondere die Ausführungen der vorstehenden Abschnitte im Zusammenhang mit hybriden Wortbildungen.

Charakteristika“ im Zusammenhang mit dem Punkt „Ambinguität“ vertiefend dargestellt.

Vielfalt orthographischer Bezeichnungen

Gerade die bereits in einem vorstehenden Abschnitt beschriebene hybride Beschaffenheit der medizinischen Sprache führt in der Praxis zur häufigen Koexistenz unterschiedlicher orthographischer Bezeichnungen.
Insbesondere durch die „Eindeutschung“ lateinischer, griechischer oder englischer Wörter ergeben sich Effekte, die sich wie folgt anhand eines charakteristischen Beispiels aufzeigen lassen:
Die „Eindeutschung“ lateinischer Termini führt in vielen Fällen zur Ersetzung von Beugungssuffixen sowie zur Substitution des lateinischen „c“ durch ein deutsches „k“ (Beispiel: Ulcus ventriculi vs. Magenulkus oder Collum uteri vs. Uteruskollum).
Die Existenz unterschiedlicher orthographischer Ausprägungen erfordert bei der Analyse medizinischer Texte einen flexiblen Umgang mit Abweichungen von möglichen „Norm-Orthographien“. Diese Anforderung überschneidet sich in der Praxis stark mit Forderungen nach der Korrektur möglicher Schreibfehler.

2.2 Inhaltliche Charakteristika

Nominalsprachlicher Charakter

Wingert[44] charakterisiert die medizinische Sprache als eine Fachsprache, die eine Stellung zwischen natürlicher und formaler Sprache einnimmt. Diese Charakteristik spiegelt sich im Sprachstil medizinischer Texte wieder, der in der Regel eine im Gegensatz zur allgemeinen Sprache restriktive und vereinfachte Struktur besitzt.[45]

Schalck[46] arbeitet durch statistische Häufigkeitsanalysen deutlich heraus, daß Nomen in der ärztlichen Fach- bzw. Alltagssprache bevorzugt verwendet werden. Man kann somit von der ärztlichen Fachsprache als einer „Nominalsprache“ sprechen.

Dieser nominalsprachliche Charakter ermöglicht ein zielgerichtetes Vorgehen bei der Analyse entsprechender Texte, wobei gerade der Auswertung von Nomen folglich eine besondere Rolle zukommt.

---

44 Vgl. [WINGERT 89].
45 Vgl. [SCHERRER 90] sowie [NANGLE 94].
46 Vgl. [SCHALCK 74].

Ambinguität

Bei der Analyse medizinischer Texte tritt häufig der Fall auf, daß ein sprachlicher Term nicht eindeutig auf ein gedankliches Konzept abgebildet werden kann. Verantwortlich hierfür sind im wesentlichen Homographie und Homonymie:[47]

Homographie

Von Homographie spricht man, wenn verschiedene Terme zwar unterschiedlich ausgesprochen, aber gleich geschrieben werden. Als Beispiel lassen sich „Tenor“ und „Tenor“ anführen. Während solche Homographe in der gesprochenen Sprache anhand des unterschiedlichen Klanges unterschieden werden können, entfallen derartige Unterscheidungskriterien bei der Betrachtung geschriebener Texte, so daß zusätzliche Analysekriterien herangezogen werden müssen.

Homonymie

Von Homonymie spricht man, wenn der gleiche Term für verschiedene gedankliche Konzepte stehen kann. Als Beispiel lässt sich der Term „Bruch“ anführen, der gedanklich für „Hernie“ oder „Fraktur“ stehen kann.

Sollen medizinische Texte analytisch erschlossen werden, so ist es in der Regel notwendig, die im jeweiligen Text vorhandenen Terme eindeutig auf Konzepte abzubilden. Liegt hierbei einer der vorstehend beschriebenen Fälle von Ambinguität vor, so werden Zusatzinfomationen benötigt, um den Kontext zu erschließen, in dem sich der abzubildende Term befindet. Der Versuch, mit Hilfe kontextueller Zusatzinformationen die Ambinguität sprachlicher Terme zu überwinden, wird als Disambiguierung bezeichnet.
Die beschriebene Notwendigkeit zur Disambiguierung tritt insbesondere auch im Zusammenhang mit Abkürzungen auf. Die medizinische Sprache ist, wie bereits beschrieben, in der Regel auf Komprimierung bedacht, so daß in der Praxis zahlreiche Abkürzungen verwendet werden.

[47] In der medizinischen Fachsprache spielen darüber hinaus auch „Homophone“ (verschiedene Terme werden trotz unterschiedlicher Schreibweise gleich ausgesprochen) eine Rolle. Wenn man bei der Analyse von medizinischen Texten aber von korrekter Orthographie ausgeht, spielen „Homophone“ an dieser Stelle keine Rolle.

Findet man beispielsweise in einem medizinischen Text die Abkürzung „HWI“, so ist eine Disambiguierung notwendig, da mit der Abkürzung grundsätzlich „Hinterwandinfarkt“ oder „Harnwegsinfekt“ gemeint sein könnte.
An dieser Stelle wird unmittelbar deutlich, daß der Umgang mit Ambinguität einen wichtigen Problemkomplex im Zusammenhang mit der Analyse medizinischer Texte darstellt.

Synonymität

Synonymität bedeutet, daß ein bestimmtes Konzept sprachlich vielfältig kodiert werden kann. Im Rahmen der vorliegenden Arbeit soll weiterhin folgende Unterscheidung getroffen werden, die auf der unterschiedlichen „Reichweite“ synonymer Strukturen beruht und die eine für die Analyse medizinischer Texte sinnvolle Gliederung induziert:

- *Synonymität auf Wortebene*
- *Synonymität auf Phrasenebene*

Synonymität auf Wortebene

Nachfolgend werden zunächst Phänomene beschrieben, bei denen Synonymität auf die Grenzen eines einzelnen Wortes beschränkt ist. Folgende Darstellung soll dies verdeutlichen:

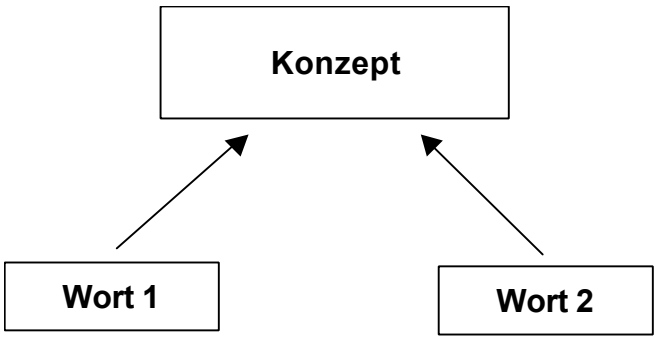

Abbildung 2: Synonymität auf Wortebene

Varianten in der Schreibung

Durch orthographische Variationen entstehen unterschiedliche Worte, die für ein identisches Konzept stehen. Als Beispiele lassen sich die Variationen „Äther“ und „Ether“ anführen.

Lexikalisch bedingte Synonyme

Lexikalisch bedingte Synonyme sind die „klassischen“ Beispiele für Synonymität. In derartigen Fällen haben mehrere mitunter völlig verschiede Worte eine identische Bedeutung. Als Beispiel ist hier das Wortpaar „Mumps“ und „Ziegenpeter“ zu nennen.

Synonymität auf Phrasenebene

In den folgenden Abschnitten wird nunmehr der Begriff der Synonymität auf Strukturen erweitert, die die Wortgrenzen überschreiten. Folgende Darstellung soll diese auf Phrasen ausgedehnte „Reichweite“ verdeutlichen:

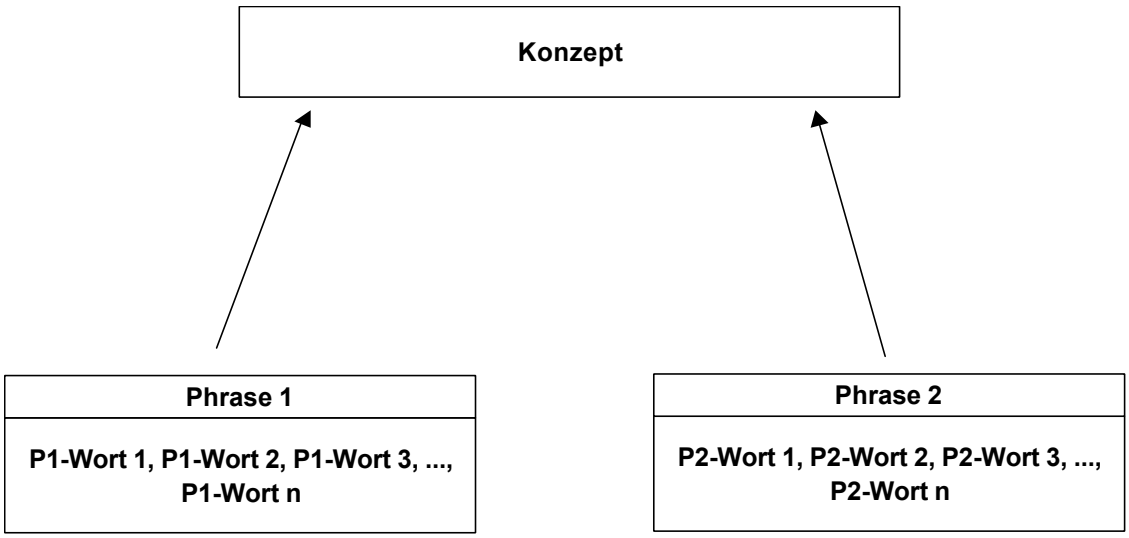

Abbildung 3: Synonymität auf Phrasenebene

Komposita

Komposita stellen die einfachste Form von Strukturen dar, bei denen das Phänomen der Synonymität über Wortgrenzen hinaus auftritt. Nachdem in einem vorstehenden Absatz die Bedeutung von Komposita im Rahmen medizinischer Texte herausgearbeitet wurde, sollen nunmehr die mit der Synonymität verbundenen inhaltlichen Aspekte aufgezeigt werden. Folgendes Beispiel soll die Zusammenhänge und Konsequenzen für die Ziele der vorliegenden Arbeit anhand nachstehender Diagnosebegriffe aufzeigen:

1. Magenkrebs
2. Krebs des Magens

In beiden vorstehenden Zeilen wird derselbe Sachverhalt beschrieben. Beide Zeilen sind somit als Synonyme zu betrachten. Um die Übereinstimmung derartiger Synonyme aufzudecken, können in der Praxis Verfahren zur Zerlegung von Komposita herangezogen werden. Derartige Verfahren werden üblicherweise vor der eigentlichen Textanalyse eingesetzt und prüfen, ob ein Wort sinnvoll in Komponenten zerlegt werden kann und ob die Kombination der Einzelkomponenten tatsächlich synonym zum ursprünglichen Kompositum ist.
Bei der Durchführung einer derartigen Komponentenzerlegung ergeben sich zwei Problemkomplexe:

- Zahlreiche Komposita sind nicht einfache Konkatenationen von Einzelworten[48].
- Das Zerlegungsmuster bzw. die Granulierung der Zerlegung eines Kompositums ist nicht immer eindeutig.

Die Frage nach dem Zerlegungsmuster bzw. der Granulierung der Zerlegung eines Kompositums steht stets im Spannungsfeld von Redundanz und medizinischer Sinnhaftigkeit. Dieses Spannungsfeld lässt sich an folgendem Beispiel leicht erläutern:

[48] Siehe hierzu insbesondere auch die vorstehenden Ausführungen bezüglich Vokalschwund, etc..

Kompositum 1: Magenschleimhautkrebs

- Zerlegung 1: Magenschleimhaut – Krebs
- Zerlegung 2: Magen – Schleimhaut – Krebs

Kompositum 2: Mageneingangskrebs

- Zerlegung 1: Mageneingang – Krebs
- Zerlegung 2: Magen – Eingang - Krebs

Für die Zerlegung beider Komposita stehen jeweils zwei Möglichkeiten zur Verfügung, da beide Komposita aus jeweils drei grundsätzlich sinnvollen Einzelworten bestehen.

Folgt man bei Kompositum 1 der Zerlegung 1, so steht Schleimhaut nicht als eigenständiger Suchbegriff zur Verfügung. Magenschleimhaut muß als eigener Begriff in eine Datenbasis aufgenommen werden, obwohl die Begriffe „Magen“ und „Schleimhaut“ möglicherweise bereits vorhanden sind und eine Kombination eben dieser beiden Begriffe synonym zum Begriff „Magenschleimhaut“ ist.

Folgt man bei Kompositum 1 der Zerlegung 2, so stehen nach der Zerlegung drei medizinisch relevante Begriffe zur Verfügung und somit für weitere Bearbeitungsschritte bereit.

Bei Kompositum 2 ergibt sich ein etwas anderes Bild: Folgt man bei Kompositum 2 der Zerlegung 2, so stehen nach der Zerlegung ebenfalls drei Begriffe zur Verfügung. An dieser Stelle stellt sich aber die Frage nach der medizinischen Relevanz des Begriffes „Eingang“. Für Kompositum 2 wäre somit Zerlegung 1 die möglicherweise geeignetere Alternative.

Vorstehende Ausführungen zeigen, daß Komposita in der medizinischen Sprache eine große Rolle spielen und daß der Umgang mit diesen bei der Analyse derartiger Texte von besonderer Bedeutung ist.

Syntaktisch, semantisch und pragmatisch bedingte Paraphrasen

Syntaktisch, semantisch und pragmatisch bedingte Paraphrasen stellen die höchste Abstraktionsebene synonymer Strukturen dar. An dieser Stelle sollte man sich vor Augen halten, daß man in einer natürliches Sprache in der Regel zahlreiche verschiedene Möglichkeiten hat, einen bestimmten Sachverhalt darzustellen. Es leuchtet somit unmittelbar ein, daß dies auch für medizinische Texte gilt. Folgendes Beispiel soll dies verdeutlichen:

Aussage1:

„Die Mukosa wird lymphomonozytär infiltriert."

Aussage 2:

„Die Schleimhaut weist Infiltrate durch Lymphozyten und Monozyten auf."

Man erkennt leicht, daß beide Aussagen synonym hinsichtlich Ihrer Bedeutung „als Ganzes" sind. Gleichwohl ist diese Synonymität aber nicht durch einfache Wortzuordnungen festzustellen, d. h. es gelingt nicht, jedem Wort der einen Aussage ein korrespondierendes der anderen zuzuordnen.

An dieser Stelle wird deutlich, daß das Phänomen der syntaktisch, semantisch und pragmatisch bedingten Paraphrasen keinesfalls auf Wortebene behandelt werden kann. Ein analytischer Zugang muß auf einer Ebene mit höherem Abstraktionsgrad erfolgen. Gerade dieser analytische Zugang spielt eine große Rolle im Rahmen der Konstruktion der Datenbasis für XDIAG.

3. Konzeptionelle Grundlagen leitbegrifforientierter automatischer Diagnosekodierung

Nachdem in den vorstehenden Abschnitten automatische Diagnosekodierung in Form eines heuristischen Ansatzes theoretisch eingeführt und durch die Beschreibung ausgewählter kodierrelevanter Aspekte medizinischer Texte in einen größeren Bedeutungszusammenhang gestellt wurde, sollen nunmehr die Grundlagen von XDIAG erarbeitet werden.
Eine Vorstellung der konzeptionellen Grundlagen, die die theoretische Basis für die verwendete Analysestrategie repräsentieren, bildet den Anfang. Hierauf folgt eine Diskussion der gewählten Stammdatenbasis.

3.1 Automatische Diagnosekodierung nach *Schalck*

In den folgenden Abschnitten soll ein Verfahren vorgestellt werden, das von Herrn *Detlef Schalck* auf Basis der Erfahrungen im Rahmen seiner langjährigen Tätigkeit im Bereich medizinische Dokumentation an der Deutschen Klinik für Diagnostik in Wiesbaden (DKD) vorgeschlagen wurde. Eine explizite Veröffentlichung des Verfahrens liegt zwar nicht vor, doch lassen sich aus den unveröffentlichten Unterlagen und Notizen *Schalcks* seine Konzepte und Ideen wenigstens vom Grundsatz her zuverlässig herausarbeiten. Besonders wichtig in diesem Zusammenhang ist auch die Tatsache, daß viele Ideen und Konzepte im Rahmen gemeinsamer Tätigkeiten mit Fachkollegen erarbeitet wurden und einer dieser Fachkollegen, *Herr Prof. em. Dr. med. Wolfgang Giere*, bei Unklarheiten und Verständnisschwierigkeiten im Rahmen der Erstellung der vorliegenden Arbeit stets unterstützend herangezogen werden konnte.

Die Beschäftigung mit den Ideen und Konzepten *Schalcks* als Basis einer umfangreicheren Ausarbeitung erscheint grundsätzlich insbesondere aus folgenden Gründen sinnvoll:

- *Schalck* brachte eine jahrzehntelange Erfahrung im Bereich medizinische Dokumentation in seine Konzepte und Ideen ein.
- *Schalck* kam als ein ehemaliger Mitarbeiter von *Prof. em. Dr. med. W. Giere* bereits früh mit innovativer Informationstechnologie im medizinischen Umfeld in Berührung.

- Die Nähe *Schalcks* zur medizinischen Praxis führte zu einem permanenten Abgleich seiner zunächst theoretischen Ansätze mit der medizinischen Alltagswelt.

Nachdem nunmehr die Motivation für die Beschäftigung mit den Konzepten und Ideen *Schalcks* verdeutlicht wurde, soll im folgenden das für die Zwecke der vorliegenden Arbeit relevante Gedankengut herausgearbeitet werden.

Im Mittelpunkt des von *Schalck* vorgeschlagenen Verfahrens zur automatischen Diagnosekodierung steht das Konzept der „Leitbegrifforientierung“, das sehr eng mit dem Konzept der „Einzelwortorientierung“ verknüpft ist. Beide Konzepte werden wegen ihrer grundlegenden Bedeutung im folgenden definiert und in den Kontext der vorliegenden Arbeit eingeführt.

In den folgenden Abschnitten ist besonders zu beachten, daß die dargestellten Ideen und Konzepte *Schalcks* Teil einer größeren Ausarbeitung im Zusammenhang mit der Entwicklung und Konzeption grundlegender Verfahren zur Verarbeitung medizinischer Texte sind. Eine vollständige Darstellung aller Konzepte und Ansätze würde den Rahmen der vorliegenden Arbeit sprengen, so daß an dieser Stelle eine Extraktion der für die automatische Diagnosekodierung relevanten Inhalte erfolgt.

3.1.1 Das Konzept der „Leitbegrifforientierung“

Im Rahmen seiner Ausarbeitungen entwickelte *Schalck*[49] unter anderem ein Modell zur automatischen Diagnosekodierung, das er selbst als „Eklipse“-Konzept bezeichnet. Dieses Modell findet sich in den Aufzeichnungen *Schalcks* in unterschiedlichen Entwicklungs- bzw. Ausbaustufen. Im folgenden wird nur die aktuellste und somit wahrscheinlich vollständigste Darstellung vom 18.08.1999 berücksichtigt.

Bereits anhand der Bezeichnung „Eklipse“-Konzept sowie auf der Basis zahlreicher unterschiedlicher Notizen von *Schalck* wird klar, daß er hiermit ein rechnergestütztes Kodiersystem für Diagnosen bezeichnet, das im Hintergrund bestehender Dokumentationssysteme arbeitet und dort den bereits im Rahmen des BAIK-Modells beschriebenen Zusatznutzen erbringt.

---

[49] Vgl. [SCHALCK 02].

Das Basiskonzept und somit der Kern des Verfahrens ist der Vergleich von aus dem zu analysierenden Diagnosetext aufbereiteten Wort-Mustern[50] mit in einer Datenbank abgelegten Wort-Mustern, denen auf Basis einer Kodiersystematik im Vorfeld ein Diagnosekode zugeordnet wurde. Für einen Diagnosetext soll somit derjenige Kode automatisch vergeben werden, bei dem eine maximale Übereinstimmung der verglichenen String-Muster festzustellen ist. Eine syntaktische Analyse der Eingabedaten erfolgt nicht.[51]

Von besonderer Bedeutung im Rahmen des beschriebenen Ansatzes ist das Konzept des „Leitbegriffs", das im folgenden kurz erläutert werden soll. Besonders interessant in diesem Zusammenhang ist die Feststellung, daß das Konzept des Leitbegriffs in den Überlegungen *Schalcks* deutlichen Modifikationen unterworfen ist. Man erkennt einen „Reifungsprozeß", der auf eine intensive konzeptionelle Bearbeitung des Themas durch *Schalck* schließen läßt. Am Ende dieser intensiven Bearbeitung steht folgende Charakterisierung des Konzeptes „Leitbegriff":

Leitbegriffe sind Determinanten von Krankheitsbeschreibungen

Zimmer[52] definiert „Begriff" als „die Bedeutungsvorstellung, die ein Wort im Geist hervorruft". Man kann somit vermuten, daß „Leitbegriff" für *Schalck* diejenigen Worte charakterisiert, die im Bewußtsein des medizinisch geschulten Hörers unmittelbar und rudimentär mit bestimmten Krankheitszuständen verbunden sind. Somit sind Leitbegriffe eben solche Teile medizinischer Texte, die mental besonders eng mit bestimmten Krankheiten assoziiert sind.[53]

---

[50] In der Sprache der Informatik könnte man an dieser Stelle auch von einem stringorientierten Mustervergleich sprechen, wobei der Begriff „Muster" eine Menge von „Strings" kennzeichnet.

[51] Vgl. hierzu auch die Ausführungen von *Zips* und *Giere* [ZIPS 83], in denen eine Klartextverarbeitung auf Basis der Erkenntnisse von *Schalck* anderen syntaxorientierten Verfahren gegenübergestellt wird.

[52] Vgl. [ZIMMER 99].

[53] Vgl. zum Leitbegriff-Konzept die Ausführungen von *De Bruijn* [DEBRUIJN 98], der von „content words" als besonders sinntragend für einen bestimmten medizinischen Text spricht.

Leitbegriffe determinieren somit den direkten gedanklichen Zugang[54] zu Krankheiten – ein Ansatz der offensichtlich ausgezeichnet gerade in das Umfeld der automatischen Diagnosekodierung paßt, da es hierbei eben auf einen effizienten und effektiven Zugang im Sinne einer möglichst exakten Kodeermittlung ankommt.[55]

Weiterhin läßt sich vermuten, daß *Schalck* mit dem Konzept „Leitbegriff" auch eine Reduktion von Worten auf Vorzugsbegriffe anstrebte, da verschiedene synonyme Worte beim Hörer gleiche Bedeutungsvorstellungen hervorrufen und somit den gleichen Begriff repräsentieren.

Im Sinne einer praktischen Umsetzung des Konzeptes wurden von *Schalck* folgende Arten von Leitbegriffen vorgeschlagen:[56] [57]

Typ „LS"

Einzelwörter, bei denen Zusätze die Kodevergabe nicht beeinflussen.

Beispiel: Ketonurie (ICD-10: R82.4)

Typ „LK"

Krankheitsbezeichnungen, bei denen durch Modifikatoren die Kodevergabe beeinflußt wird.

Beispiel: Fraktur, Finger (ICD-10: S62.60)
Fraktur, Fuß (ICD-10: S92.9)

Typ „LT"

Topographiebezeichnungen, die zur Kodevergabe einen oder mehrere Modifikatoren erfordern.

Beispiel: Akutes Abdomen (ICD-10: R10.0)

---

[54] Vgl. hierzu die Erläuterungen in [ICD10 99].

[55] Vgl. hierzu insbesondere die Ausführungen von *Röttger* [RÖTTGER 73a] im Zusammenhang mit der Struktur eines „Basis-Diagnosesatzes". *Röttger* führt aus, daß in vielen Fällen bereits die Mitteilung eines einfachen Diagnosebegriffes als zwischenärztliche Mitteilung sinnvoll sein kann. Diagnosebegriffe bilden den „Kern" der in einem Diagnosesatz formulierten Information.

[56] Vgl. hierzu auch das von *Franz et al.* [FRANZ 00] durchgeführte „Ranking", das unmittelbar diagnosebezogenen Informationen bei der Ermittlung von Diagnosekodes erste Priorität einräumt.

[57] Aus Gründen der leichteren Nachvollziehbarkeit erfolgt die Darstellung der Beispiele anhand der aktuellen ICD-10 [ICD10 03a].

Typ „LH“

Homonyme, die eine Sonderbehandlung erfordern.

Beispiel: Bruch (Fraktur oder Hernie?)

Schalck gliedert auch die verschiedenen Arten von Modifikatoren sehr detailliert. Die Zielsetzung dieser Gliederung ist aber, wie bereits angedeutet, ganz deutlich eine erweiterte Textanalyse mit dem Ziel, medizinische Texte inhaltlich qualifiziert zu erschließen. Da gerade dieser Ansatz aber nicht Gegenstand der vorliegenden Arbeit ist, wird er, neben anderen Aspekten der erweiterten Textanalyse, an dieser Stelle nicht vertiefend dargelegt.

3.1.2 Der idealtypische Ablauf der leitbegrifforientierten automatischen Diagnosekodierung

Die Umsetzung des vorstehend beschriebenen Basiskonzepts im Rahmen eines vollständigen Verfahrens zur automatischen Diagnosekodierung läßt sich nach *Schalck* wie folgt kurz und idealtypisch umreißen:

I. Vorbereitung

I.a Erstellung eines Einzelwortkataloges

Ein Einzelwortkatalog soll alle Wörter des jeweiligen Schlüsselsystems ergänzt um Wörter aus dem praktischen klinischen Sprachgebrauch (praktischer diagnostischer Wortschatz) aufnehmen. Hierbei soll eine Kennzeichnung erfolgen, ob das jeweilige Wort ein Leitbegriff, ein Modifikator oder ein „Stopwort“ (für die Kodierung irrelevant und somit entbehrlich[58]) ist. Man beachte insbesondere, daß hier von Einzelwörtern gesprochen wird und somit implizit eine Vorverarbeitung vorhandener Komposita vorausgesetzt wird. Weiterhin fordert *Schalck* eine Abbildung von Wort-Beugungsformen auf die entsprechenden Grundformen („Normalisierung“) sowie eine Auflösung von Abkürzungen. Zur Realisierung vorgenannter Funktionen werden von *Schalck* Ersetzungslisten vorgeschlagen.

[58] Vgl. [MOORE 94a] sowie das Konzept der „function words“ von *De Bruijn* [DEBRUIJN 98].

I.b Erstellung eines Schlüsselkataloges

In einem Schlüsselkatalog sollen die String-Muster zusammen mit den zugeordneten Kodes gemäß der jeweils relevanten Schlüsselsystematik (aus Gründen der Aktualität und Nachvollziehbarkeit soll im folgenden der ICD-10 beispielhaft als Basis des Schlüsselkataloges dienen) abgelegt werden. *Schalck* weist hierbei insbesondere darauf hin, daß dieser Schlüsselkatalog so klein wie möglich zu halten ist. Aus diesem Grunde sollten Wiederholungen identischer String-Kode-Kombinationen durch unterschiedliche Wortstellungen sowie durch unterschiedliche Deklinations- bzw. Konjugationsformen vermieden werden[59].

Zur Vermeidung von Wiederholungen identischer String-Kode-Kombinationen schlägt *Schalck* eine alphabetische Sortierung der normalisierten Modifikatoren mit nachfolgender Verkettung mit dem Leitbegriff vor.

Folgendes Beispiel[60] soll dies verdeutlichen:

Kode: C30.1 Text: Bösartige Neubildung am Mittelohr

Verarbeitungsschritte zur Erstellung des Schlüsselkatalogs:

„bösartige“
-> nach Normalisierung: „bösartig“

„Neubildung“
-> nach Normalisierung: „Neubildung“

„am“
-> als Stopwort zu entfernen!

„Mittelohr“
-> nach Normalisierung: „Mittelohr“

Es ergibt sich somit nach alphabetischer Sortierung[61] und Konkatenation folgender Eintrag in den Schlüsselkatalog:

Kode: C30.1 Text: neubildungbösartigmittelohr

---

[59] Vgl. hierzu insbesondere die Ausführungen im Zusammenhang mit der Erstellung des Einzelwortkataloges.
[60] Vgl. [ICD10 03a].
[61] Diese von Schalck konzipierte alphabetische Sortierung ist als Äquivalent der bei Xmed durchgeführten Standardisierung der Eingangstexte zu betrachten (vgl. [LUZ 97]).

II. Automatische Ermittlung der Kodes

Im folgenden soll nunmehr die von *Schalck* entworfene automatische Ermittlung der Kodes idealtypisch skizziert werden:

Zunächst erfolgt die textuelle Aufarbeitung. Hierzu werden in einem ersten Schritt alle irrelevanten Wörter (Stopworte) aus dem Text entfernt. Anschließend werden Rechtschreibfehler korrigiert sowie Abkürzungen aufgelöst. *Schalck* läßt an dieser Stelle den genauen Ablauf der Fehlerkorrektur offen, man kann aber vermuten, daß auch an dieser Stelle der Einsatz von Ersetzungslisten angedacht war. Unklare Abkürzungen sollen unter Berücksichtigung des jeweiligen medizinischen Fachgebiets aufgelöst werden.

Hierauf müssen in dem zu kodierenden Abschnitt die vorkommenden und nunmehr auch im Einzelwortkatalog vorhandenen Wörter gekennzeichnet werden (Leitbegriff, Modifikator, etc.). Auf diese Weise werden die vorhandenen Leitbegriffe hervorgehoben und als Determinanten der jeweils zu vergebenden Kodes herausgearbeitet.

Anschließend muß mit Hilfe des Schlüsselkataloges ermittelt werden, welche für den jeweiligen Leitbegriff kodierrelevanten Modifikatoren im vorliegenden Diagnosetext vorhanden sind.

Aus den relevanten Modifikatoren wird durch alphabetische Sortierung und anschließende Konkatenation mit dem zugehörigen Leitbegriff das zu vergleichende String-Muster generiert. Mit diesem Muster wird nunmehr im Schlüsselkatalog der am besten geeignete Kode gesucht.

Während bei *Schalck* somit der Mustervergleich anhand sortierter und somit quasi normalisierter String-Muster erfolgt, findet man in diesem Zusammenhang in der Literatur zahlreiche Ansätze[62], die zur Bestimmung eines geeigneten Kodes komplexe Vektormodelle[63] nutzen.

---

[62] Vgl. hierzu insbesondere [SURJAN 01], [DEBRUIJN 97], [DEBRUIJN 98] und [HASMAN 01].

[63] Vektormodelle repräsentieren im vorliegenden Zusammenhang Algorithmen aus dem Bereich der linearen Algebra, die die Übereinstimmung zweier String-Muster durch Ermittlung eines Abstandsmaßes quantifizieren. Die Suche nach einem geeigneten Kode für ein vorliegendes String-Muster würde somit durch die Suche nach einem bereits kodierten String-Muster mit geringem Abstand realisiert.

An dieser Stelle findet man bei *Schalck* leider keinen Hinweis darauf, in welcher Weise unklare Entscheidungssituationen aufzulösen sind:
Das Konzept des Leitbegriffs induziert, daß ein Wort in einem Kodierzusammenhang als Leitbegriff und in einem anderen als Modifikator zu betrachten ist.
Folgendes Beispiel soll dies verdeutlichen:

Diagnose: „akutes Abdomen“

In diesem Falle ist „Abdomen“ ein Leitbegriff

Diagnose: „Stichverletzung im Abdomen“

In diesem Falle ist „Abdomen“ ein Modifikator

Somit ist es in der Praxis schwierig, zu entscheiden, ob z.B. das Wort „Abdomen“ im vorliegenden Zusammenhang ein Leitbegriff ist oder nicht.
Weiterhin erscheint es grundsätzlich problematisch, beim Vorkommen mehrerer Leitbegriffe in einem zu kodierenden Textabschnitt die richtige Zuordnung der Modifikatoren zu den jeweiligen Leitbegriffen und somit zu möglicherweise vorhandenen Diagnosezusammenhängen zu treffen.

3.2 Der ICD-10-Diagnosen-Thesaurus (IDT) als geeignete Stammdatenbasis

Nachdem nunmehr das Grundkonzept der automatischen leitbegrifforientierten Diagnosekodierung nach *Schalck* eingeführt ist, stellt sich die Frage, auf Basis welcher Stammdaten ein solches Verfahren realisiert werden kann. Vor allen anderen Dingen ist an dieser Stelle zu klären, auf welche Weise die zentralen Datenkataloge (Einzelwortkatalog sowie Schlüsselkatalog) in der Praxis mit Leben gefüllt werden können. Da diese Frage von grundlegender Bedeutung ist, erfolgt deren Diskussion vor der Betrachtung konkreter Realisierungsansätze.[64]

3.2.1 Inhaltsaspekte des IDT

Der IDT ist eine Sammlung von Krankheitsbegriffen im deutschen Sprachraum, verschlüsselt nach ICD-10-SGB-V, Version 2.0. Ziel der Herausgabe des IDT war und ist es, ein Werk unterstützend zur ICD-10-SGB-V-Version bereitzustellen, um auf diese Weise die Verschlüsselung von Diagnosen in der täglichen Praxis zu erleichtern. Der

Schwerpunkt des IDT liegt somit auf der Berücksichtigung des täglichen Sprachgebrauchs in der Praxis tätiger Mediziner.

Bei der Erarbeitung des IDT wurden bis zur Version 4.0 folgende Begriffssammlungen berücksichtigt [DIMDI 01]:

- Begriffe der ICD-10-Systematik
- Begriffe des alphabetischen Verzeichnisses der ICD-10
- Begriffe des Thesaurus der ehemaligen Arbeitsgruppe Klartextanalyse der GMDS, bis 31. März 2003 gepflegt am Zentrum der Medizinischen Informatik des Klinikums der Johann Wolfgang Goethe-Universität Frankfurt am Main
- Beiträge ärztlicher Berufsverbände mit vorwiegend Routinediagnosen aus den Fachgebieten Allgemeinmedizin, Anästhesiologie, Augenheilkunde, Kinder- und Jugendheilkunde sowie Mund-Kiefer-Gesichtschirurgie
- „Göttinger Diagnosen“ (ICD-9-/ICD-10-Diagnosensammlung) von Dr. Bernd Graubner (anfänglich für die Georg-August-Universität Göttingen zusammengestellt)
- Begriffe der ICD-9-Diagnosensammlung der IMS GmbH, Frankfurt am Main (Institut für Medizinische Statistik)
- Begriffe aus dem von Dr. Reinhart Köhler erarbeiteten Diagnosenverzeichnis des gynäkologischen EDV-Anwenderkreises GYNAMED
- Begriffe, die 1997 im ICD-10-Modellversuch der Kassenärztlichen Vereinigungen Niedersachsen und Sachsen-Anhalt gesammelt worden waren
- Begriffe aus dem ADT-Panel des ZI, das auf anonymisierten Abrechnungsdaten niedergelassener Ärzte aus verschiedenen Fachrichtungen beruht

Anhand vorstehender Auflistung erkennt man leicht, daß der IDT eine umfassende Darstellung des in der ärztlichen Fachsprache genutzten Wortschatzes repräsentiert.

---

[64] Als Basis für die nachstehenden Ausführungen dient der IDT Version 4.0 [DIMDI 01].

Dieser enge Praxisbezug zeichnet die Daten des IDT bereits ohne Betrachtung möglicher Strukturaspekte als für die Bearbeitung medizinischer Texte relevante Datenbasis aus. Weiterhin kann auf Grund der breitbasigen Datengewinnung von einer hohen Vollständigkeit[65] der Datenmenge des IDT ausgegangen werden.
Gerade diese Vollständigkeit wird im Rahmen eines aktuellen Datenpflegeprojektes weiter ausgebaut:

Beim DIMDI, Köln wird zur Zeit ein Datenpflegeprojekt bearbeitet, dessen Ziel die **vollständige** Integration des Alphabetischen Verzeichnisses der ICD-10 (WHO-Fassung) in den IDT ist. Man erkennt leicht, daß dieser Schritt einen wichtigen Meilenstein im Rahmen der Bearbeitung des IDT markiert. Nach Abschluß dieses Projektes werden die Stammdaten des IDT die Diagnosetexte des gesamten Alphabetischen Verzeichnisses der ICD-10 (WHO-Version) beinhalten.[66]

Während man die zunächst praxisorientiert durchgeführte Sammlung von Begriffen als „Bottom-UP-Ansatz“ bezeichnen kann, repräsentieren die Datenmengen des Systematischen Verzeichnisses sowie des Alphabetischen Verzeichnisses der ICD-10 (WHO-Version) einen „Top-Down-Ansatz“. Die Begriffsmenge des IDT lässt sich somit als ein Resultat eines Gegenstrom-Mechanismus bezeichnen, der für das bereits erwähnte Maximum an Vollständigkeit und Relevanz bürgt.

---

65 Vgl. [CIMINO 98].

66 Das Projekt wurde im Dezember 2004 abgeschlossen. Das Alphabetische Verzeichnis der ICD-10 ist nunmehr vollständig in den IDT integriert; IDT und Alphabetisches Verzeichnis sind zu einem einheitlichen Werk verschmolzen.

Dieser Gegenstrom-Mechanismus lässt sich an Anlehnung an vorstehende Ausführungen wie folgt veranschaulichen:

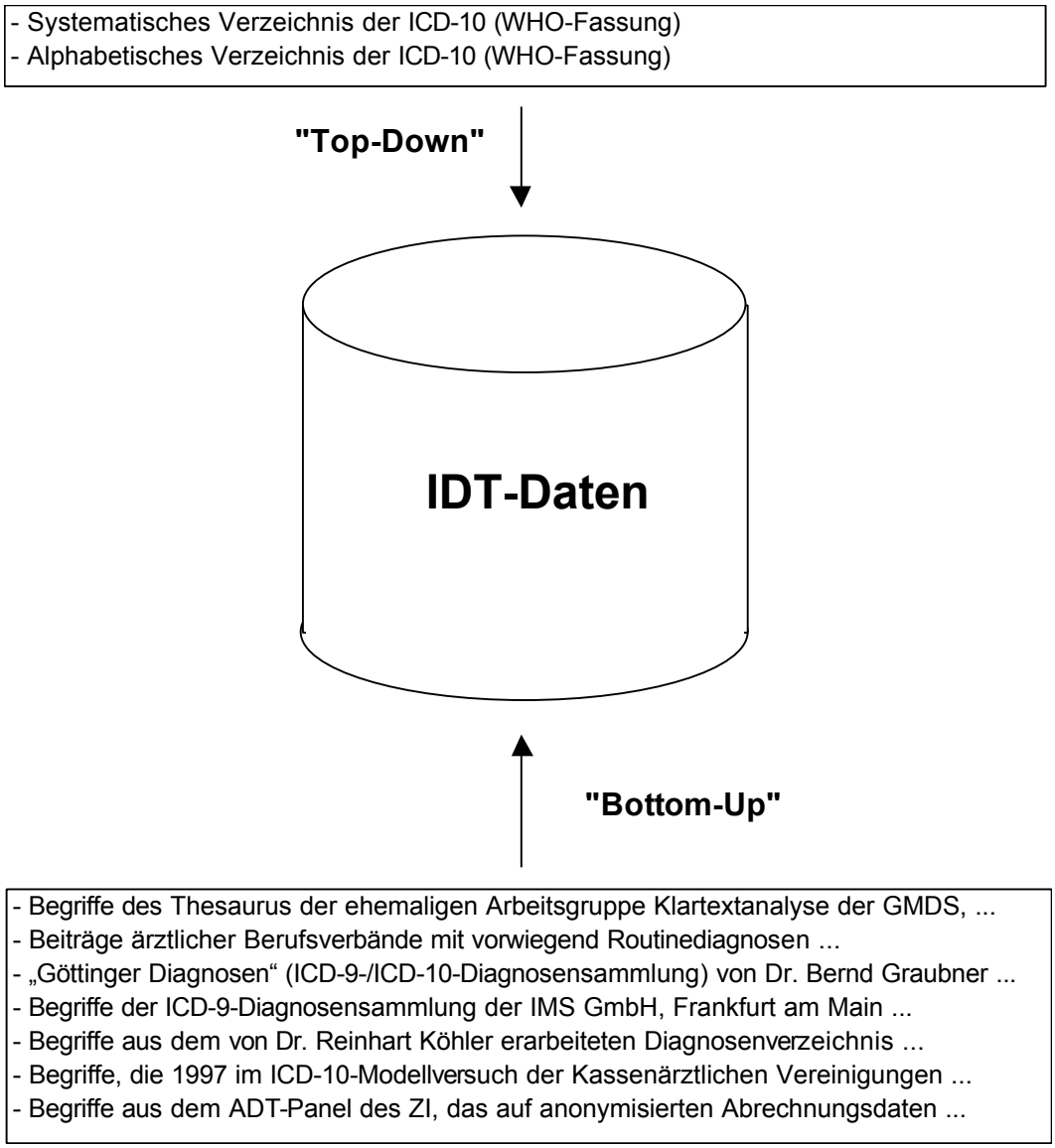

Abbildung 4: Der Datenbestand des IDT

Zusammenfassend lässt sich nunmehr festhalten, daß der Datenbestand des IDT die von *Schalck* zur leitbegrifforientierten Diagnosekodierung geforderte Menge an Eingangsbegriffen in idealer Weise repräsentiert.

Weiterhin lässt sich festhalten, daß der IDT auch als Basis zur Erstellung des Schlüsselkataloges hervorragend geeignet ist. Bereits in Abschnitt 2.2 der vorliegenden Arbeit wurde im Zusammenhang mit Synonymität auf Phrasenebene aufgezeigt, daß die medizinische Sprache unter anderem geprägt ist von syntaktisch, semantisch und

pragmatisch bedingten Paraphrasen. Im Rahmen der Analyse medizinischer Texte mit dem Ziel einer Ermittlung von Diagnosekodes bedeutet dies, daß eine Vielzahl unterschiedlichster Texte zu identischen Kodes führen, da diese verschiedenen Texte eben gerade Paraphrasen desselben medizinischen Sachverhaltes sind.[67] Die Daten des IDT können somit im Rahmen einer automatischen Kodierung helfen, in der medizinischen Praxis verwendete sprachliche Spielarten einer Analyse zugänglich zu machen. Auf diese Weise können auch auf den ersten Blick ungewöhnliche Umschreibungen einer Diagnose sauber analysiert und kodiert werden.
An dieser Stelle ist nochmals die permanente Datenpflege des IDT hervorzuheben, da auch die Analyse von Synonymität auf Phrasenebene erheblich von einer „lebendigen Datenbasis“ profitiert.

3.2.2 Strukturaspekte des IDT

Nachdem im vorstehenden Abschnitt die für automatische Verarbeitung medizinischer Texte relevanten inhaltlichen Aspekte des IDT herausgearbeitet wurden, soll nunmehr geprüft werden, ob sich strukturelle Eigenschaften finden lassen, die ebenfalls für eine entsprechende Eignung der IDT-Daten sprechen.

Der IDT wird aktuell in zwei Versionen publiziert:

Softwareversion
Die Softwareversion enthält ausformulierte Diagnosetexte mit den dazugehörigen Schlüsselnummern.

Buchversion
Die Buchversion enthält die Diagnosetexte mit Schlüsselnummern in permutierter und alphabetisch sortierter Form, so daß über die entstehenden Sucheinträge ein Zugang über die verschiedenen Komponenten der Diagnosetexte möglich ist. Dieser Zugang lässt sich leicht an folgendem Beispiel aufzeigen:

Softwareversion:
Magenkarzinom; C16.9

Sucheinträge der Buchversion:
Magen, Karzinom; C16.9
Karzinom, Magen; C16.9

[67] An dieser Stelle sei nochmals auf das Beispiel in Abschnitt 2.2. hingewiesen.

Bei der Arbeit mit der Buchversion wird der Kode C16.9 somit gefunden, wenn die Kodesuche mit dem Stichwort Magen *oder* mit dem Stichwort Karzinom beginnt.
Zur Vermeidung einer doppelten und somit redundanten Datenhaltung für beide Versionen sind beide Formen (Softwareversion und Buchversion) eines jeden IDT-Eintrages miteinander verknüpft. Die Buchversion wird aus der Softwareversion mit Hilfe semantischer Informationen erzeugt. Diese Informationen legen fest, auf welche Weise der Eintrag gegebenenfalls aufzutrennen und zu permutieren ist.
Die vorstehend beschriebenen semantischen Informationen werden durch eine SGML-Auszeichnung[68] der Diagnosetexte realisiert.

Möchte man sich mit Blick auf ein zu realisierendes Verfahren zur automatischen Diagnosekodierung einen ersten Überblick über grundsätzliche strukturelle Eigenschaften des IDT verschaffen, so erscheint es sinnvoll, die vorhandenen SGML-Auszeichnungen zunächst in Form eines Überblicks auszuwerten. Diese zunächst einfachen Auswertungen können natürlich im Rahmen der durch die SGML-Struktur vorgegebenen Möglichkeiten verfeinert werden, falls dies für die Lösung von Detailproblemen bei der Realisierung des Prototypen notwendig sein sollte.[69]

Am Zentrum der Medizinischen Informatik, Frankfurt wurde im Februar 2002 eine Strukturanalyse des IDT Version 3.1 durchgeführt, um auf diese Weise einen Überblick über die Möglichkeiten der Verwendung der Daten im Rahmen eines Verfahrens zur automatischen Diagnosekodierung auf Basis der Konzepte von *Schalck* treffen zu können. Untersucht wurden hierbei die SGML-Auszeichnungen („Tags“) der einzelnen IDT-Einträge, da diese die Art und somit die Semantik der Komponenten der einzelnen Einträge abbilden und auf diese Weise bestimmte strukturelle Analysen erlauben.

---

[68] SGML steht für „Standardized Generalized Markup Language“ (ISO 8879 (1986)); SGML ist eine „Metasprache“, die es erlaubt, Auszeichnungssprachen zu definieren. Auszeichnungssprachen wiederum erlauben es, in einem Textdokument durch vereinbarte Zeichen („Tags“) semantische Strukturen zu definieren und diese somit einer automatisierten Verarbeitung zugänglich zu machen. Eine auf Basis von SGML durchgeführte Auszeichnung eines Textes repräsentiert somit semantische Zusatzinformationen über die eigentlichen Inhalt einzelner Worte hinaus. Für einen guten Überblick über SGML siehe beispielsweise [GOLDFARB 90].

[69] Vgl. hierzu insbesondere die Ausführungen in Kapitel 2.1 im Zusammenhang mit der Frage einer geeigneten Zerlegung von Komposita.

Um zu möglichst übersichtlichen Ergebnissen zu kommen, wurden Computerprogramme entwickelt, die im wesentlichen die Aufgabe zu erfüllen hatten, die SGML-Auszeichnung einer jeden IDT-Zeile auf die für die Strukturanalyse relevanten „Tags“ zu reduzieren.

Mit Blick auf das Ziel der Analyse waren dies folgende „Tags“:

<NP>; <A>; <S>; <U>; <P>; <F>[70]

Die vorstehenden „Tags“ haben hierbei folgende Bedeutung, wobei ein Tag in der Mehrheit der Fälle ein einzelnes Wort repräsentiert.[71]

<NP>	Nominalphrase[72]
<S>	Substantiv
<A>	Adjektiv
<U>	Unselbständiger Wortbestandteil[73]
<P>	Präposition

Tabelle 1: Auszeichnungs-Tags des IDT

[70] An dieser Stelle sei darauf hingewiesen, daß natürlich auch die gemäß SGML-Konvention schließenden „Tags“ berücksichtigt wurden; aus Gründen der Übersichtlichkeit werden diese hier aber nicht näher beschrieben, da sie leicht aus den öffnenden „Tags“ durch Voranstellen des Zeichens „/“ abgeleitet werden können.

[71] Eine Besonderheit findet man beim Auftreten von „Termini Technici“: Derartige feststehende medizinische Fachbegriffe werden wie ein Wort behandelt, auch wenn es sich um einen Mehrwortbegriff handelt. Beispiel: „Arteria renalis“. Durch dieses Vorgehen wird verhindert, daß eben diese „Termini Technici“ bei der Erstellung der Buchversion getrennt und unsachgemäß permutiert werden.

[72] Nominalphrasen sind sprachliche Einheiten, die aus einer Verkettung von Wörtern nach gewissen grammatikalischen Regeln gebildet werden. Üblicherweise werden hierbei Substantive und Adjektive verknüpft. Nominalphrasen enthalten jedoch keine Verben [ZAISS 02]. Im IDT bilden Nominalphrasen die wichtigste hierarchische Strukturkomponente.

[73] Unselbständige Wortbestandteile sind Wörter, die nur im Zusammenhang mit einem zugehörigen Substantiv sinnvoll verwendet werden können. Ein typisches Beispiel hierfür ist „Korsakow-Syndrom“, wobei „Korsakow“ der unselbständige Wortbestandteil ist.

Die IDT-Einträge wurden abschließend anhand der erhaltenen „Tag"-Muster sortiert. Es hat sich hierbei gezeigt, daß 80% der IDT-Einträge strukturell durch 8 Muster abgebildet werden.

Das Ergebnis läßt sich wie folgt zusammenfassen:[74]

Muster	**Anteil**	**Bemerkung** **Beispiel**
<NP><S><S></NP>	25%	Zwei Substantive *Karzinom, Magen*
<NP><S></NP>	19%	Ein Substantiv *Embolie*
<NP><S><A></NP>	15%	Ein Substantiv – ein Adjektiv *Hypersensitiv, Blase*
<NP><U><S></NP>	8%	Ein unselbständiger Wortbestandteil – ein Substantiv *Korsakow-, Syndrom*
<NP><A><S><S></NP>	5%	Zwei Substantive – ein Adjektiv *Akute, Fraktur, Tibia*
<NP><NP><S></NP><P><NP><S></NP></NP>	3%	Zwei Substantive durch eine Präposition verbunden *Infektion durch Herpesvirus*
<NP><A><A><S></NP>	2%	Ein Substantiv – zwei Adjektive *Akute, myeloische, Leukämie*
<NP><A><NP><U><S></NP></NP>	2%	Ein unselbständiger Wortbestandteil und ein Substantiv als Einheit – ein Adjektiv *Akutes, Korsakow-, Syndrom*

Tabelle 2: Verteilung der Auszeichnungs-Tags des IDT

Die vorstehende Tabelle zeigt, daß der IDT von Einträgen geprägt ist, bei denen Substantive entweder alleine stehen oder von anderen Substantiven, Adjektiven oder unselbständigen Wortbestandteilen modifiziert werden. Die Strukturelemente sind hierbei stets Einzelworte.[75] [76]

---

[74] Um eine bessere Lesbarkeit zu erreichen und gleichzeitig die vorhandene Struktur transparenter herauszuarbeiten, wurden für den „Tag" „<np>" auch die jeweils schließenden „Tags" aufgetragen.

[75] Auf die „Termini Technici" als Ausnahmen wurde bereits hingewiesen.

[76] Vgl. in diesem Zusammenhang auch das von *Röttger* [RÖTTGER 73a] beschriebene Konzept des „Basis-Diagnosesatzes". Ein Basis-Diagnosesatz ist eben durch die Beantwortung der drei Grundfragen „Was? – Wie? – Wo?" gekennzeichnet.

Berücksichtigt man die vorstehend aufgezeigte überragende strukturelle Bedeutung der Substantive im IDT und hält man sich vor Augen, daß Substantive gleichfalls zentrale Träger des „Leitbegriff"-Konzeptes von *Schalck* sind, so läßt sich vermuten, daß die Daten des IDT durchaus zu eben diesem Konzept passen:

Für die Einträge, bei denen nur ein Substantiv existiert, ergibt sich der Leitbegriff automatisch; für die Einträge, bei denen mehrere Substantive vorliegen, ist eine differenzierte Betrachtung notwendig, die in einem späteren Abschnitt der vorliegenden Arbeit eingehend erläutert wird.

Hält man sich abschließend nochmals vor Augen, daß der IDT speziell auch den praktischen klinischen Sprachschatz repräsentiert, so wird klar, daß bei Referenzierung der IDT-Daten die Anzahl der für einen Kode notwendigen Einzelworte in 80% der Fälle kleiner oder gleich drei ist, wobei die Kodierung durch mindestens ein enthaltenes Substantiv determiniert wird. Diese Fakten sprechen einerseits für die Machbarkeit und andererseits für die Effizienz einer einzelwortorientierten leitbegrifforientierten automatischen Diagnosekodierung auf Basis der IDT-Daten.

4. Konzeption von „XDIAG“

An dieser Stelle soll nunmehr die Konzeption des Prototypen beschrieben werden, der im Rahmen der vorliegenden Arbeit zur Realisierung und Evaluation der von *Schalck*[77] vorgeschlagenen leitbegrifforientierten automatischen Diagnosekodierung („XDIAG“) entwickelt wurde.

Anhand der einführenden Definitionen läßt sich leicht ableiten, daß es sich bei eben dieser Konzeption um die abstrakte Spezifikation eines Systems von Computerprogrammen zusammen mit einer geeigneten Stammdatenbasis handelt.

Im Rahmen einer Einführung wird der Prototyp[78] zunächst im Überblick dargestellt. Dieser Überblick soll helfen, die grundsätzlichen Bearbeitungsabläufe zu verstehen und in das auf Basis der Konzepte *Schalcks* in einem vorstehenden Abschnitt der vorliegenden Arbeit dargelegte idealtypische Schema einzuordnen.

In einem zweiten Schritt werden im Sinne einer schrittweisen Verfeinerung die einzelnen realisierten Teilfunktionen dargelegt. Da es sich stets um Programme[79] handelt, kann man an dieser Stelle auch von einer Beschreibung der Einzelmodule sprechen. Bei eben dieser Beschreibung werden neben den realisierten Funktionen natürlich auch die hierbei benötigten Daten eingeführt.

In einem abschließenden Abschnitt werden schließlich die Struktur sowie die Erstellung der vorstehend erwähnten Daten[80] erläutert. Diese Reihenfolge erscheint sinnvoll, da nur bei Berücksichtigung des vollständigen funktionellen Rahmens die Datenbasis wirklich adäquat verstanden und diskutiert werden kann.

---

77 Vgl. [SCHALCK 02].

78 In allen nachfolgenden Abschnitten der vorliegenden Arbeit soll im Sinne einer besseren Lesbarkeit der Ausführungen der Begriff „Prototyp“ stets die entwickelte „Konzeption des Prototypen“ bezeichnen.

79 Die Konzeption der Programme orientiert sich an anerkannten und bewährten Regeln der Informatik. Vgl. hierzu beispielsweise [MEHLHORN 88] sowie [AHO 74].

80 In allen nachfolgenden Abschnitten der vorliegenden Arbeit soll die Gesamtmenge aller zur Funktion des Prototypen benötigten Daten stets als „Datenbasis“ bezeichnet werden.

4.1 Der Prototyp im Überblick

Nachstehendes Schaubild erlaubt einen einfachen Überblick über die grundsätzlichen Bearbeitungsschritte. Im Sinne einer schrittweisen Verfeinerung wird zunächst ein Gesamtüberblick gegeben, der in einem nächsten Schritt feiner aufgegliedert wird.

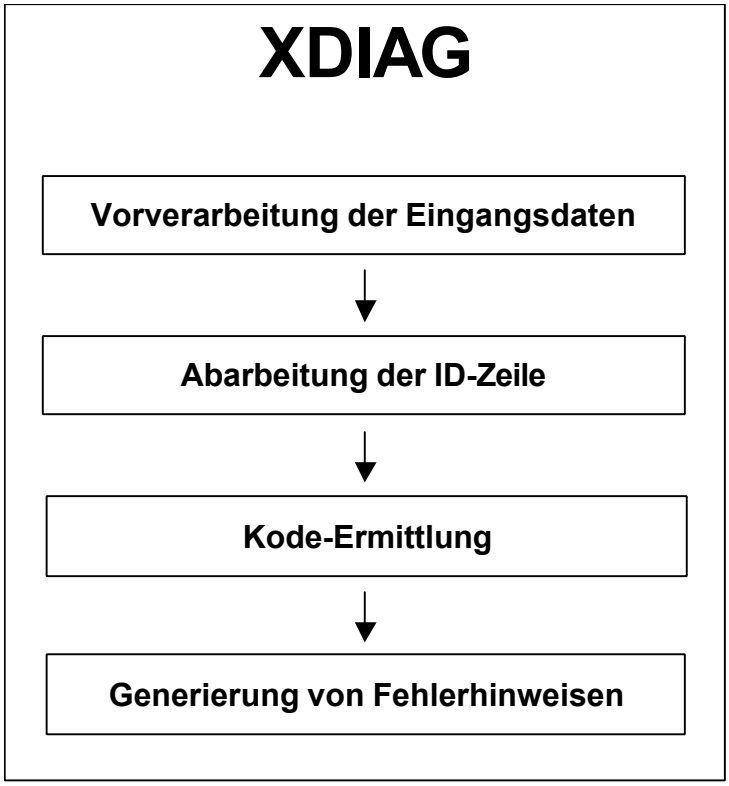

Abbildung 5: Die Bearbeitungsschritte von XDIAG

4.2 Konzeption der Einzelkomponenten

In Erweiterung der vorstehenden Übersicht soll nunmehr, wie bereits beschrieben, eine detaillierte Beschreibung der Einzelkomponenten vorgenommen werden.

4.2.1 Die Vorverarbeitung der Eingangsdaten

Der erste Verfahrensabschnitt dient der Aufbereitung der in Form medizinischer Texte bereitgestellten Eingangsdaten. Man kann davon ausgehen, daß gerade dieser Verfahrensabschnitt bzw. die hierbei realisierten Funktionen erhebliche Auswirkungen auf die erzielbaren Ergebnisse haben.

Im Sinne einer besseren Übersicht sowie im Sinne einer leichteren Nachvollziehbarkeit der Ausführungen erfolgt zunächst eine Darstellung der Einzelkomponenten in Form eines Schaubildes:

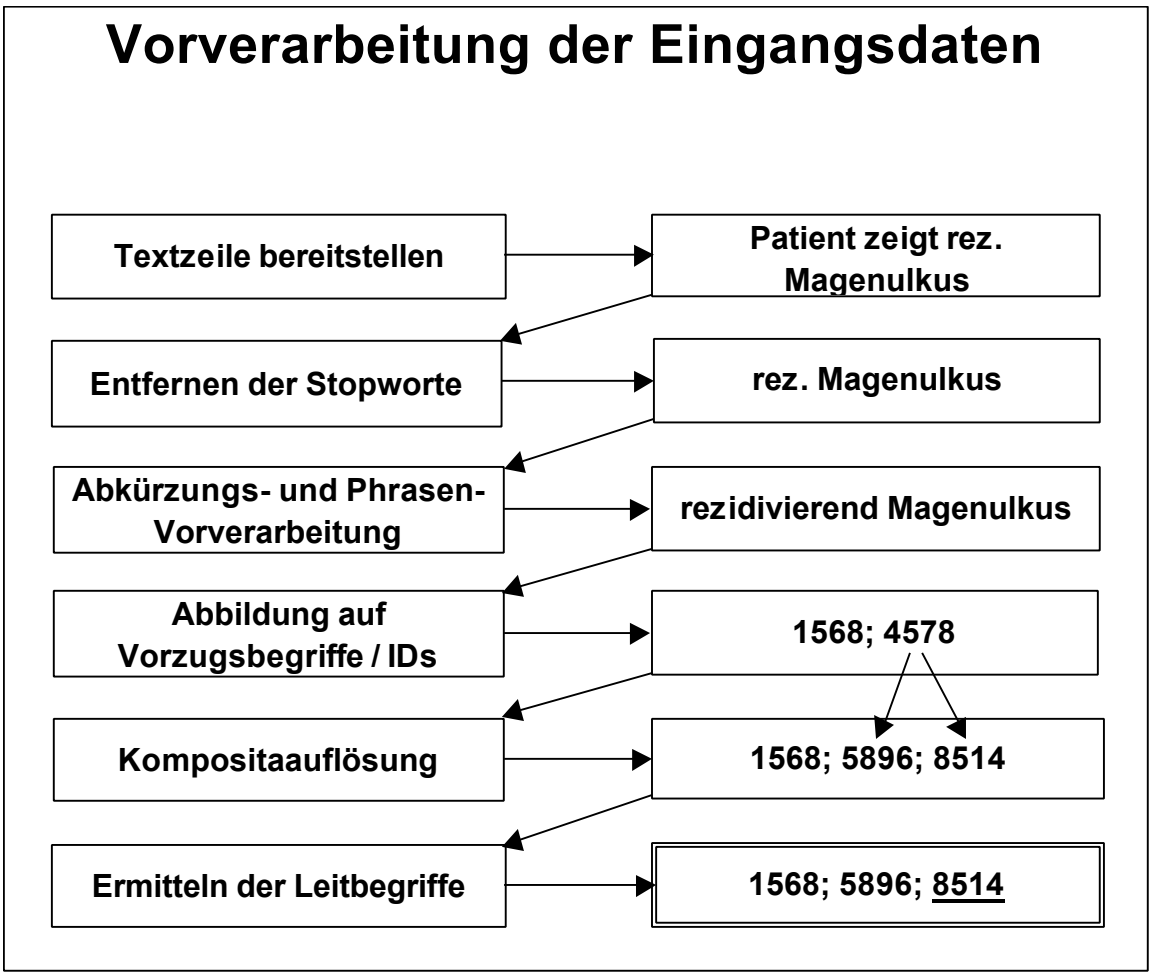

Abbildung 6: Die Vorverarbeitung der Eingangsdaten

4.2.1.1 Bereitstellung einer Textzeile

Um eine flexible Integrierbarkeit des entwickelten Prototypen zu erreichen, muß eine geeignete Spezifikation der Eingangsdaten vorgenommen werden, da sonst kein sinnvoller Aufsetzpunkt zur qualifizierten Diskussion der Konzeption der nachfolgenden Verarbeitungsfunktionen gefunden werden kann.

Für die Zwecke der vorliegenden Arbeit soll an dieser Stelle folgende Spezifikation der Eingangsdaten vorgenommen werden:

Die von XDIAG zu verarbeitenden Texte müssen feldweise einlesbar bzw. verarbeitbar bereitgestellt werden. Eine im Quellsystem vorhandene und möglicherweise semantisch bedeutsame Satzstruktur[81] (beispielsweise durch Satzzeichen repräsentiert) soll bei der notwendigen feldweisen Übergabe der Daten an den Prototypen erhalten bleiben.

Die Funktionalität des Prototypen an dieser Stelle reduziert sich also auf die feldweise Zwischenspeicherung der Eingangsdaten, wobei die jeweiligen Feldinhalte in der Regel durch Satzzeichen determinierte Zeilen repräsentieren.

Die zu diesem Zeitpunkt entstandene und für die weiteren Schritte bereitgestellte Struktur soll im folgenden als „Eingabefeld“ bezeichnet werden. Die in einem solchen Eingabefeld enthaltenen Substrukturen sollen als Worte bezeichnet werden, unabhängig davon, ob es sich beispielsweise um Abkürzungen oder römische Zahlen handelt. Wichtig an dieser Stelle ist nur die grundsätzliche strukturelle Gliederung eines Eingabefeldes in Worte.

4.2.1.2 Entfernen der Stopworte

Stopworte werden gemäß einer Stopwortliste zunächst automatisch aus dem Eingabefeld entfernt. Auf diese Weise wird der in den nachfolgenden Schritten zu bearbeitende Text kompakter, so daß die Verarbeitungsgeschwindigkeit deutlich zunimmt.
Die automatische Entfernung wird im Rahmen des Prototypen wie folgt realisiert:
Das Eingabefeld wird Wort für Wort mit einer hinterlegten Stopwortliste verglichen. Wird ein Wort als Stopwort identifiziert, so wird es aus dem Eingabefeld entfernt.

4.2.1.3 Abkürzungs- und Phrasen-Vorverarbeitung

In einem einführenden Abschnitt der vorliegenden Arbeit wurde das von *Schalck* vorgeschlagene Verfahren zur leitbegrifforientierten Diagnosekodierung als grundsätzlich „einzelwortorientiert“ charakterisiert. Es treten in der Praxis aber zahlreiche Fälle auf, in denen ein Wort als Einzelwort nicht sinnvoll im Rahmen der verfolgten Zielsetzung ausgewertet werden kann.[82]

---

[81] An dieser Stelle soll besonders auf die Bedeutung der Grenzen von Sätzen für die Struktur eines Textes hingewiesen werden. Auch wenn im Rahmen der vorliegenden Arbeit kein echtes Textverständnis angestrebt wird, so sind Satzgrenzen doch als wichtige und unentbehrliche Strukturelemente zu betrachten, die bei der Bereitstellung der zu untersuchenden Texte adäquat repräsentiert werden müssen.

[82] *Moore* [MOORE 89] weist im Rahmen der automatischen Verarbeitung medizinischer Texte auf die Notwendigkeit einer effektiven Erkennung und Berücksichtigung von Mehrwortbegriffen hin.

Derartige Fälle lassen sich grob wie folgt gliedern:

- Verarbeitung von Abkürzungen
- Verarbeitung von „Termini Technici“

Verarbeitung von Abkürzungen

In Texten des medizinischen Alltags spielen Abkürzungen, wie bereits beschrieben, eine große Rolle. Bei analytischer Betrachtung stellt man fest, daß Abkürzungen in vielen Fällen nur bei Berücksichtigung zusätzlicher Informationen zuverlässig ausgewertet werden können. Als Zusatzinformationen können einerseits wortorientierte und andererseits wortübergreifende Aspekte herangezogen werden.

Als **wortorientierte** Informationen kommen beispielsweise Groß- und Kleinschreibung in Frage.

Die **wortübergreifende** Gewinnung von Zusatzinformationen läßt sich leicht an folgendem Beispiel erläutern:

„M.“

In den meisten Fällen dürfte diese Abkürzung für „Musculus“ stehen. Gleichwohl könnte hiermit aber auch „Morbus“ abgekürzt werden. Um an dieser Stelle zu einer vernünftigen Abschätzung[83] der Bedeutung der aufgefundenen Abkürzung zu kommen, ist es sinnvoll, zunächst zu untersuchen, ob hinter der Abkürzung ein Wort bzw. eine Wortgruppe steht, die einen Muskel oder eine Krankheit bezeichnen könnte.

Verarbeitung von „Termini Technici“

Medizinische Fachsprache ist, wie bereits beschrieben, in ganz erheblichem Maße durch die Verwendung von Fachbegriffen geprägt. Im Rahmen einer automatischen Diagnosekodierung stellen derartige Fachbegriffe eine besondere Herausforderung dar. Bestehen solche „Termini Technici“ nämlich aus mehreren Worten, so führt eine konsequent einzelwortorientierte Vorgehensweise zu erheblichen Problemen bei der Analyse entsprechender Texte.

[83] Es erscheint an dieser Stelle sinnvoll, von Abschätzung zu sprechen, da die automatische Aufarbeitung von Abkürzungen, wie bereits in Kapitel 2.1 erwähnt, stets mit einer gewissen Unsicherheit behaftet ist.

Dieses Phänomen soll nachfolgend anhand eines Beispiels erläutert werden:

Sehnerv = Nervus opticus = Hirnnerv II

Man erkennt leicht, daß die Analyse des Wortes „Sehnerv“ sich einfach gestaltet. Betrachtet man nunmehr die Analyse des Wortpaares „Nervus opticus“, so liegt auf der Hand, daß bei einer einzelwortorientierten Vorgehensweise die Zusammengehörigkeit beider Worte nicht als Determinante der weiteren Textanalyse automatisch erschlossen werden kann. Dies gilt insbesondere für das Wortpaar „Hirnnerv II“. Gerade bei einem Wort, das eine Zahl repräsentiert, ist es besonders schwer, automatisch den jeweils analytisch korrekten Zugang zu finden.

Anhand der vorstehenden Beispiele wird leicht klar, warum bei der Bearbeitung von Abkürzungen bzw. „Mehrwort-Termini-Technici“[84] durch XDIAG die Berücksichtigung der jeweiligen Wortumfelder in einer frühen Phase der Analyse erfolgen muß.
Um darüber hinaus für den weiteren Ablauf der automatischen Diagnosekodierung das einzelwortorientierte Vorgehen im Sinne einer möglichst konsistenten Realisierung erhalten zu können, ist die Vorverarbeitung der Abkürzungen und Phrasen im Rahmen von XDIAG wie folgt realisiert:

Das Eingabefeld wird nach dem Prinzip des „Longest-Match“[85] nach Textbestandteilen durchsucht, die, wie beschrieben, vorverarbeitet werden müssen. Der „Longest-Match“-Ansatz stellt hierbei sicher, daß stets soviel Kontext wie möglich, d.h. soviel Text des Eingabefeldes wie möglich, zur Vorverarbeitung genutzt wird. Diese Vorgehensweise läßt ein Maximum an Präzision erwarten.
Die Vergleichsdaten für den vorstehend beschriebenen „Longest-Match“-Ansatz liefert ein Abkürzungs- und Phrasenthesaurus. Die für die Vorverarbeitung ausgewählten Textbestandteile werden mit Hilfe des jeweils relevanten Thesauruseintrags durch ein Wort oder mehrere Worte ersetzt. Diese Ersetzung stellt für den weiteren Verfahrensablauf eine sinnvolle und effektive einzelwortbasierte Analyse sicher.

---

[84] Im Sinne einer besseren Lesbarkeit der vorliegenden Arbeit sollen derartige „Mehrwort-Termini-Technici“ im folgenden als „Phrasen“ bezeichnet werden.

[85] „Longest-Match“ charakterisiert in diesem Zusammenhang die Strategie beim Vergleich des Eingabefeldes mit dem Abkürzungs- und Phrasenthesaurus: Lassen sich beim wortübergreifenden Vergleich mehrere übereinstimmende Strukturen finden, so wählt man diejenige mit der größten Länge („longest“) der zu ersetzenden Struktur aus.

An dieser Stelle sei noch auf eine zusätzliche Erweiterungsmöglichkeit hingewiesen:
Ist die fachliche Ausrichtung des Autors des zu kodierenden Textes bekannt, können, wie bereits erwähnt, auch derartige Informationen ergänzend zur Auflösung von Abkürzungen herangezogen werden.

4.2.1.4 Abbildung auf Vorzugsbegriffe mit Schreibfehlerkorrektur

Das zu diesem Zeitpunkt bereits erheblich vorverarbeitete Eingabefeld wird nunmehr wortweise auf Vorzugsbegriffe abgebildet. Der Ablauf gestaltet sich wie folgt:
Jedes Wort wird mit den Einträgen eines entsprechend gestalteten Thesaurus verglichen. Wird das gerade zu bearbeitende Wort im Thesaurus aufgefunden, so werden der Vorzugsbegriff des entsprechenden Wortes sowie die „ID“[86] eben dieses Vorzugsbegriffes aus dem Thesaurus ausgelesen. Am Ende dieses Verarbeitungsschrittes ist idealerweise der gesamte Diagnosesatz in IDs übersetzt. Das Eingabefeld wird somit durch eine Menge von IDs repräsentiert. Die Weiterverarbeitung erfolgt nunmehr abgelöst vom eigentlichen Text auf einer abstrakten Ebene. Im Sinne einer besseren Lesbarkeit soll im folgenden die in einem Arbeitsschritt abzuarbeitende Menge von IDs als „Eingabezeile“ bezeichnet werden.
Ein „Fuzzy-Algorithmus“[87] sorgt an dieser Stelle dafür, daß Schreibfehler größtenteils toleriert werden, d.h. daß auch ein falsch geschriebenes Wort in vielen Fällen auf den richtigen Vorzugsbegriff abgebildet werden kann. In Zusammenhang mit dieser schreibfehlertoleranten Vorgehensweise muß aber ganz deutlich auf die Grenzen dieses Verfahrens hingewiesen werden:
Eine zu restriktive Konfiguration der Fehlerkorrektur verhindert das Auffinden eines Vorzugsbegriffes – eine zu tolerante Fehlerkorrektur bildet möglicherweise auf falsche Vorzugsbegriffe ab.
Bereits an dieser Stelle soll vorab auf eine wichtige Anforderung an die entsprechende Datenbasis hingewiesen werden, die sich aus dem vorstehend geschilderten Verarbeitungsschritt ergibt:
In der Datenbasis müssen alle relevanten Konjugations- bzw. Deklinationsformen eines Wortes vorhanden sein, um die Abbildung aller in der Praxis möglicher Formen auf die

---

[86] Mit „ID“ soll im folgenden eine Indexnummer zur systemweit eindeutigen Kennzeichnung eines Vorzugsbegriffes bezeichnet werden.

[87] Der „Fuzzy-Algorithmus“ (vgl. [RAPP 97]) realisiert eine Trigramm-basierte Ähnlichkeitssuche, die im beschriebenen Kontext zu einer gewissen Schreibfehlertoleranz im Rahmen der Abbildung von Einzelworten der Diagnosetexte auf Vorzugsbegriffe führt.

entsprechenden Vorzugsbegriffe sicher vornehmen zu können. Derartige Varianten können zwar möglicherweise auch im Rahmen des „Fuzzy-Algorithmus“ erkannt werden, doch erkennt man leicht, daß in solchen Fällen die Präzision der Abbildung auf Vorzugsbegriffe und somit das Ergebnis der automatischen Diagnosekodierung leiden könnte.

4.2.1.5 Kompositaauflösung

Die bereits beschriebenen im Zusammenhang mit der Verarbeitung von Komposita auftretenden Probleme werden in XDIAG durch eine Vorab-Auflösung bekannter Komposita gelöst. Der Gedanke, Daten bereitzustellen, die eine Zerlegung von Komposita in ihre Bestandteile erlauben, wird bereits 1973 von *Schalck et al.*[88] formuliert.
Eben diese Vorab-Auflösung im Rahmen des entwickelten Prototypen soll im folgenden beschrieben werden:[89]

Das Eingangsfeld wird ID für ID mit den Einträgen eines Komposita-Thesaurus verglichen. Liegt für eine ID ein Thesaurus-Eintrag vor, so wird das entsprechende Einzelwort als Kompositum erkannt und durch die ID-Menge seiner Bestandteile ersetzt.

Die vorstehend beschriebene Auflösung der Komposita hat den Vorteil, daß Redundanzen bei der Datenhaltung vermieden werden. Diese Feststellung läßt sich leicht anhand folgender Überlegungen belegen:

- Nimmt man an, daß die Komponenten eines Kompositums als Solitärworte in der Eingabezeile sowie in der Datenbasis vorhanden sind[90], so läßt sich ableiten, daß bei der Analyse eines entsprechenden Diagnosetextes diese Komponenten einzeln bearbeitet und ausgewertet und somit aus der Kombination der Einträge der richtige Kode abgeleitet werden kann.
- Durch eine Auflösung der Komposita trifft dieser Ablauf auch in den Fällen zu, in denen ein Begriff in der Eingabezeile als Kompositum vorliegt: Nach der Dekomposition liegen die Komponenten der Diagnose als Solitärworte vor. Es

88 Vgl. [SCHALCK 73].
89 Vgl. [WINGERT 74], [WINGERT 85] sowie [ALTENPOHL 74].
90 Man beachte zu dieser Voraussetzung auch die entsprechenden Ausführungen im Zusammenhang mit der Erstellung der Datenbasis.

ergibt sich also für die weitere Bearbeitung und Auswertung eben gerade die Folge von Schritten, die sich auch ergeben hätte, wenn von Anfang an Solitärworte vorgelegen hätten.

Man erkennt an dieser Stelle, aus welchem Grunde der Schritt der Kompositaauflösung nach dem Schritt der Abbildung auf Vorzugsbegriffe erfolgt:
Würde man die Kompositaauflösung vor der Abbildung auf Vorzugsbegriffe durchführen, müßten die für die Auflösung notwendigen Informationen für alle in der Praxis möglichen Deklinationsformen eines jeden Kompositum hinterlegt werden. Die hieraus resultierende Datenmenge würde die vorzuhaltende Datenbasis aufblähen und die Wartung durch erhebliche Redundanzen erschweren.
Nimmt man, wie beschrieben, zunächst eine Abbildung auf Vorzugsbegriffe vor, so müssen die zur Auflösung notwendigen Informationen nur jeweils einmalig zusammen mit dem betreffenden Vorzugsbegriff abgelegt werden.

4.2.1.6 Ermittlung der Leitbegriffe

Im abschließenden Schritt der Vorverarbeitung der Eingangsdaten erfolgt die Ermittlung aller in der aktuellen Eingabezeile enthaltenen Leitbegriffe. Dieser Schritt setzt somit die „Anker“ für alle nachfolgenden Bearbeitungsschritte.
Die eigentliche Ermittlung der Leitbegriffe erfolgt durch Abgleich der einzelnen IDs der Eingabezeile mit den in Form eines Thesaurus hinterlegten Leitbegriff-Informationen. Wird eine ID als möglicher[91] Leitbegriff identifiziert, wird diese entsprechend gekennzeichnet.

Die Vorverarbeitung der Eingangsdaten ist an dieser Stelle mit folgendem Ergebnis abgeschlossen:

Das als Text übergebene Eingabefeld ist an dieser Stelle in eine durch eine ID-Menge repräsentierte Eingabezeile mit markierten möglichen Leitbegriffen überführt.

[91] Es ist wichtig, an dieser Stelle von „möglichen“ Leitbegriffen zu sprechen, da insbesondere Lokalisationen, wie bereits anhand eines Beispiels beschrieben, in bestimmten Diagnosezusammenhängen Leitbegriffe sind und in anderen nicht.

4.2.2 Die Abarbeitung der Eingabezeile

Nachdem das Eingabefeld nunmehr durch die vorstehend beschriebenen Verfahren vorverarbeitet ist und somit nur noch aus numerischen Identifikatoren besteht, die die enthaltenen Vorzugsbegriffe repräsentieren, erfolgt nun eine schrittweise (wortweise)[92] Abarbeitung mit dem Ziel, die Vorzugsbegriffe geeignet zu gruppieren, um nachfolgend einen oder mehrere Kodes ermitteln zu können.

Im Sinne einer besseren Übersicht sowie im Sinne einer leichteren Nachvollziehbarkeit der nachfolgenden Ausführungen erfolgt auch an dieser Stelle zunächst eine Darstellung der Einzelkomponenten in Form eines Schaubildes:

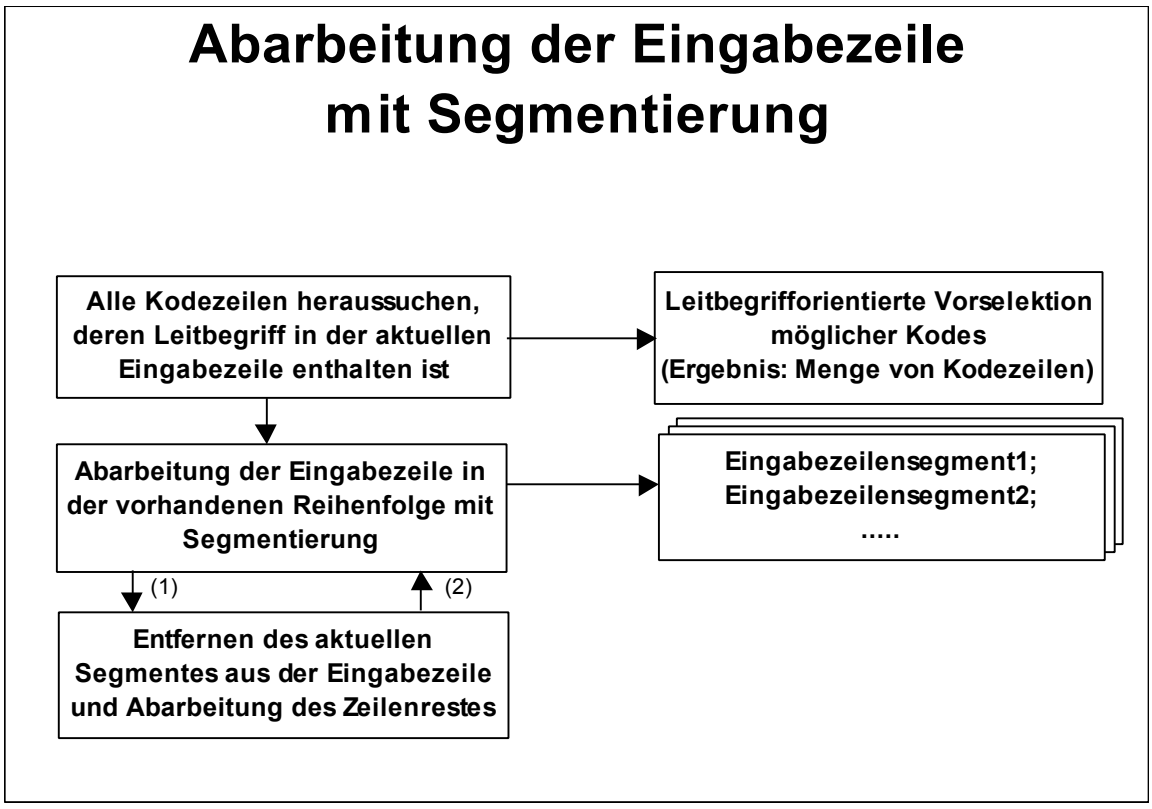

Abbildung 7: Die Abarbeitung der Eingabezeile mit Segmentierung

[92] Eigentliche müßte es an dieser Stelle „ID-weise“ heißen; mit Blick auf eine bessere Lesbarkeit wurde hier aber der Rückgriff auf Worte gewählt, da die IDs Worte eindeutig numerisch repräsentieren.

4.2.2.1 Leitbegrifforientierte Vorselektion

Der letzte Schritt der Vorverarbeitung der Eingangsdaten besteht, wie vorstehend beschrieben, aus einer Kennzeichnung der in der aktuellen Eingabezeile enthaltenen Leitbegriffe. Anhand dieser Leitbegriffe werden nun alle Diagnosekodes aus der Kodemenge herausgesucht, die einen dieser Leitbegriffe beinhalten. Es ergibt sich somit eine Vorselektion möglicher Kode-Kandidaten.
Man beachte mit Blick auf die weiteren Ausführungen, daß eben diese Kode-Kandidaten stets folgende zeilenorientierte Struktur haben:

Kode; ID_1, ID_2, ..., ID_n

Die Vorselektion und somit die Menge der Kode-Kandidaten wird also durch eine Menge vorstehend charakterisierter Zeilen repräsentiert. Das somit erreichte Zwischenergebnis läßt sich wie folgt darstellen:

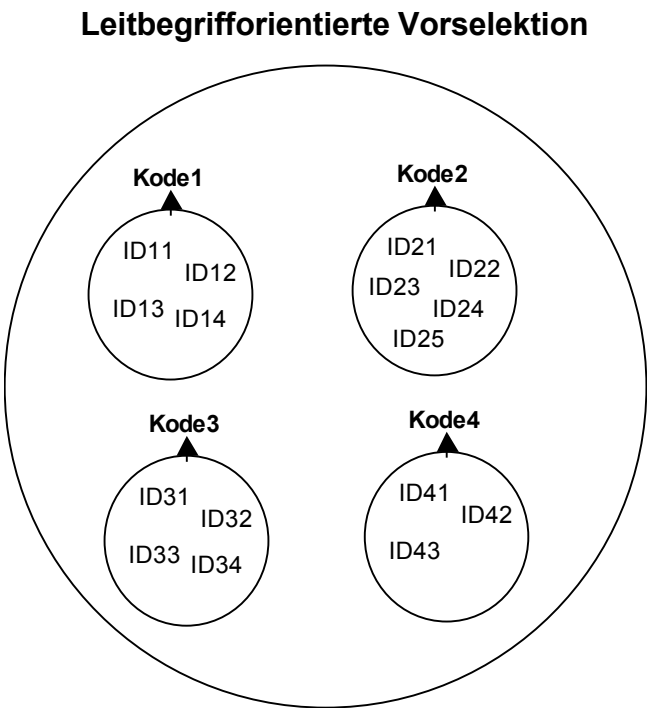

Abbildung 8: Die leitbegrifforientierte Vorselektion

Es ist unmittelbar anhand des Leitbegriff-Ansatzes abzuleiten, daß ein dem Eingabefeld schließlich zuzuweisender Kode in eben dieser Vorselektion enthalten sein muß.
Man erreicht auf diese Weise eine erhebliche Verkleinerung der in den nachfolgenden Schritten zu evaluierenden und durch Kode-Kandidaten repräsentierten Datenmenge, was in der Praxis eine deutliche Steigerung der Systemleistung erwarten läßt.

4.2.2.2 Wortweise Abarbeitung mit Segmentierung

Nach der Erstellung der Vorselektion der möglichen Kodes wird die Eingabezeile nacheinander ID für ID eingelesen und analysiert. Leitbegriffe dienen hierbei als semantische „Anker“[93], denen modifizierende Zusätze zugeordnet werden.
Ein besonders wichtiger Ansatz im Rahmen von XDIAG ist der Versuch, auch solche Eingabefelder qualifiziert zu bearbeiten, die mehrere Diagnosen beinhalten. Um eben diesen Versuch einer „Segmentierung“ besser einordnen zu können, erscheint es sinnvoll, an dieser Stelle kurz auf ausgewählte Charakteristika bestehender Systeme hinzuweisen:
Im Rahmen vorhandener Verfahren wird das Problem, daß in medizinischen Texten mehrere Diagnosen hintereinander stehen können, häufig wie folgt gelöst:

1. Man fordert die Punkt-zu-Punkt-Regel[94], d.h. man fordert die Trennung einzelner Diagnosen durch ein eindeutig identifizierbares Satzzeichen[95].

2. Man kodiert pro Diagnosezeile nur eine Diagnose; die mögliche Existenz mehrerer Diagnosen wird explizit ignoriert.[96]

Berücksichtigt man die am Zentrum der Medizinischen Informatik, Frankfurt bei der Bearbeitung vergangener Projekte im Zusammenhang mit automatischer Diagnosekodierung gemachten Erfahrungen, so wird deutlich, daß die vorstehenden Prämissen in vielen Fällen dem angelieferten Textmaterial nicht gerecht werden.

---

93 Siehe hierzu insbesondere die Ausführungen im Rahmen der Einführung des Leitbegriff-Konzeptes.
94 Vgl. [RÖTTGER 69].
95 Da dieses eindeutig identifizierbare Satzzeichen in der Regel ein Punkt ist, spricht man von der „Punkt-zu-Punkt-Regel“.
96 Vgl. [FRANZ 00].

In der Praxis bedeutet dies, daß die auf Basis eben dieser Prämissen automatisch erzielten Kodierergebnisse unvollständig sein können, was zu einer Verfälschung hieraus abgeleiteter Ergebnisse führen dürfte.

Bei der Erstellung von XDIAG wurde aus diesem Grunde ein Verfahren entwickelt, das, ebenfalls auf heuristischer Basis, die aufbereitete Textzeile bestmöglich auf Basis des Leitbegriff-Ansatzes mit Blick auf die Ermittlung von Diagnosekodes segmentiert und auf diese Weise auch die automatische Ermittlung mehrerer Kodes in einer Zeile erlaubt.

Der Ablauf der Abarbeitung der Eingabezeile mit Segmentierung läßt sich wie folgt darstellen:

Alle IDs der Eingabezeile werden nacheinander zu einer Testmenge hinzugefügt. Die der Testmenge hinzugefügten IDs werden aus der Eingabezeile entfernt. Die Testmenge wird nach Ergänzung einer jeden einzelnen ID daraufhin untersucht, ob diese Testmenge eine Teilmenge einer Kodezeile aus der Vorselektion ist.

Wenn ja, wird die Testmenge aktualisiert und die Bearbeitung schreitet zur nächsten ID der Eingabezeile fort. Alle zu diesem Zeitpunkt als „verworfen“ gekennzeichneten IDs der Eingabezeile werden aus der Eingabezeile entfernt.

Wenn nein, wird die zuletzt in die Testmenge aufgenommene ID wieder aus der Testmenge entfernt. Weiterhin wird geprüft, ob diese ID einen Leitbegriff repräsentiert.

Wenn ja, wird die bestehende Testmenge abgeschlossen und als Eingabezeilensegment zur Weiterverarbeitung bereitgestellt. Eine neue leere Testmenge wird initialisiert. In diese neue Testmenge wird die aktuelle ID als erster Leitbegriff des neuen Segments übernommen. Alle zu diesem Zeitpunkt als „verworfen“ gekennzeichneten IDs werden der aktuellen Eingabezeile (dem aktuellen Eingabezeilenrest) vorangestellt. Die Bearbeitung schreitet nunmehr mit der nächsten ID der Eingabezeile fort.

Wenn nein, wird die aktuelle ID als „verworfen“ markiert. Anschließend schreitet die Bearbeitung ohne weitere Zwischenschritte mit der nächsten ID der Eingabezeile fort.

Mit der Bearbeitung der letzten ID der Eingabezeile ergibt sich automatisch das letzte Segment einer Eingabezeile.

Das vorstehend beschriebene Verfahren nutzt das bereits beschriebene zentrale Charakteristikum[97] des leitbegrifforientierten Vorgehens:

Leitbegriffe sind Determinanten von Diagnosen und diese Determinanten eignen sich somit als Anker für die Suche nach modifizierenden Zusätzen im jeweiligen Anker-Umfeld. Wertet man somit die Umfelder bei identifizierten Ankern aus, lassen sich, bei stets vorhandener Unschärfe, mehrere Diagnosen in einer Zeile sinnvoll abgrenzen.

[97] An dieser Stelle sei nochmals auf die Definition des Leitbegriff-Ansatzes hingewiesen.

Der Ablauf der Abarbeitung der Eingabezeile mit Segmentierung läßt sich zusammenfassend anhand eines Beispiels in Form zweier Grafiken darstellen:[98]

Text der Eingabezeile:

Neben einer akuten Niereninsuffizienz findet man eine inguinale Hernie.

Vorverarbeitung

Ausgangswort	Verarbeitung	Ergebnis
neben	Stopwort entfernen	
einer	Stopwort entfernen	
akuten	abbilden auf Vorzugsbegriff	akut
Niereninsuffizienz	Kompositaauflösung	Niere
		Insuffizienz (Leitbegriff)
findet	Stopwort entfernen	
man	Stopwort entfernen	
eine	Stopwort entfernen	
inguinale	abbilden auf Vorzugsbegriff	inguinal
Hernie	abbilden auf Vorzugsbegriff	Hernie (Leitbegriff)

Eingabezeile nach Vorverarbeitung:

akut; Niere; Insuffizienez; inguinal; Hernie

Abbildung 9a: Beispiel: Die Vorverarbeitung der Eingabezeile

[98] Im Sinne einer leichteren Nachvollziehbarkeit des Beispiels wurden an Stelle der IDs die jeweiligen Vorzugsbegriffe im Klartext aufgeführt.

Wortweise Abarbeitung mit Segmentierung

aktuelles Wort	Bearbeitungsschritt	aktuelle Testmenge	aktuelle Eingabezeile
akut	"akut" zu Testmenge hinzufügen	akut	Niere Insuffizienz inguinal Hernie
	gibt es ein Element der Vorselektion, das alle IDs der aktuellen Testmenge enthält? - ja!		
Niere	"Niere" zu Testmenge hinzufügen	akut; Niere	Insuffizienz inguinal Hernie
	gibt es ein Element der Vorselektion, das alle IDs der aktuellen Testmenge enthält? - ja!		
Insuffizienz	"Insuffizienz" zu Testmenge hinzufügen	akut; Niere; Insuffizienz	inguinal Hernie
	gibt es ein Element der Vorselektion, das alle IDs der aktuellen Testmenge enthält? - ja!		
inguinal	"inguinal" zu Testmenge hinzufügen	akut; Niere; Insuffizienz; inguinal	Hernie
	gibt es ein Element der Vorselektion, das alle IDs der aktuellen Testmenge enthält? -nein!		
	"inguinal" aus der Testmenge entfernen	akut; Niere; Insuffizienz	
	"inguinal" als "verworfen" kennzeichnen		
	ist "inguinal" ein Leitbegriff? - nein!		
Hernie	"Hernie" zu Testmenge hinzufügen	akut; Niere; Insuffizienz; Hernie	(leer)
	gibt es ein Element der Vorselektion, das alle IDs der aktuellen Testmenge enthält? -nein!		
	"Hernie" aus der Testmenge entfernen	akut; Niere; Insuffizienz	
	ist "Hernie" ein Leitbegriff? - ja!		
	aktuelle Testmenge abschließen; **Ergebnis: Testmenge1: akut; Niere; Insuffizienz**		
	Neue Testmenge initialisieren und "Hernie" zu dieser zunächst leeren Testmenge hinzufügen	Hernie	
	"inguinal" (als "verworfen" gekennzeichnet) der aktuellen Eingabezeile (leer) voranstellen		inguinal
inguinal	"inguinal" zu Testmenge hinzufügen	Hernie; inguinal	(leer)
	gibt es ein Element der Vorselektion, das alle IDs der aktuellen Testmenge enthält? -ja!		
	nächstes Wort zu Testmenge hinzufügen - Eingebezeile leer - Testmenge2 abgeschlossen! - **Ergebnis: Testmenge2: Hernie; inguinal** Ende der Segmentierung - Testmengen zur Kodierung bereitstellen.		

Abbildung 9b: Beispiel: Die Segmentierung der Eingabezeile

Als Resultat des vorstehend beschriebenen Analyseschrittes erhält man eine Menge von Eingabezeilensegmenten. Jedes Eingabezeilensegment wird nunmehr zunächst einzeln für die nachfolgende Kodeermittlung bereitgestellt und in Form einer abgeschlossenen Einheit analysiert.

4.2.3 Die Kodeermittlung

Nach der beschriebenen Vorverarbeitung folgen nunmehr die Verarbeitungsschritte zur eigentlichen Kodeermittlung. Falls diese nicht gelingt, können auf Basis der während der laufenden Analyse gewonnenen Informationen qualifiziert Fehlerhinweise generiert und dem Benutzer in Form möglicher Ergänzungen übermittelt werden.

4.2.3.1 Ermittlung von Einzelkodes

Das Ergebnis des im vorstehenden Abschnittes diskutierten Verarbeitungsschrittes bilden, wie bereits erwähnt, Eingabezeilensegmente, wobei jedes dieser Segmente durch eine ungeordnete Zusammenstellung von IDs (ID-Menge) repräsentiert ist. Man arbeitet nunmehr Segment für Segment ab, mit dem Ziel, für jedes dieser Segmente einen geeigneten Diagnosekode zu ermitteln.
Zur Ermittlung eines geeigneten Kodes erfolgt ein Vergleich der in dem zu untersuchenden Segment enthaltenen IDs mit den IDs der Elemente der im vorstehenden Abschnitt beschriebenen Vorselektion. Für jedes Element der Vorselektion[99] wird untersucht, ob die Menge der dort enthaltenen IDs eine Teilmenge des aktuellen Segments ist. Alle Kodezeilen, für die diese Teilmengen-Beziehung erfüllt ist, werden zwischengespeichert und als finale Kodezeilen an den abschließenden optimierenden Verarbeitungsschritt übergeben.
Folgende Abbildungen verdeutlichen vorstehend beschriebenen mengenorientierten Analyseschritt:

[99] Man sollte sich an dieser Stelle nochmals verdeutlichen, daß die Elemente der Vorselektion jeweils durch eine Menge von IDs zusammen mit dem korrespondierenden Kode repräsentiert sind.

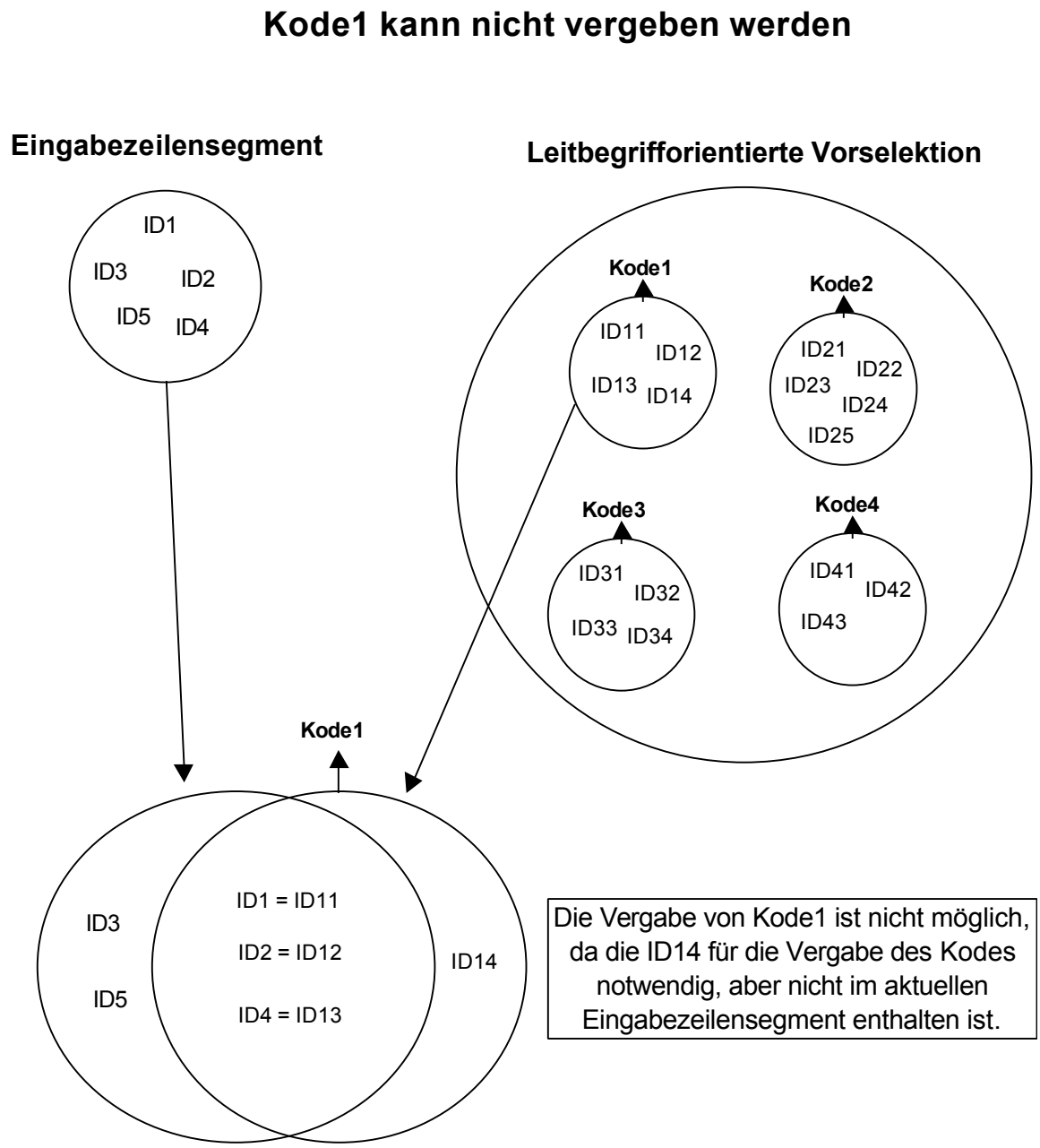

Abbildung 10: Beispiel: Kodevergabe nicht möglich

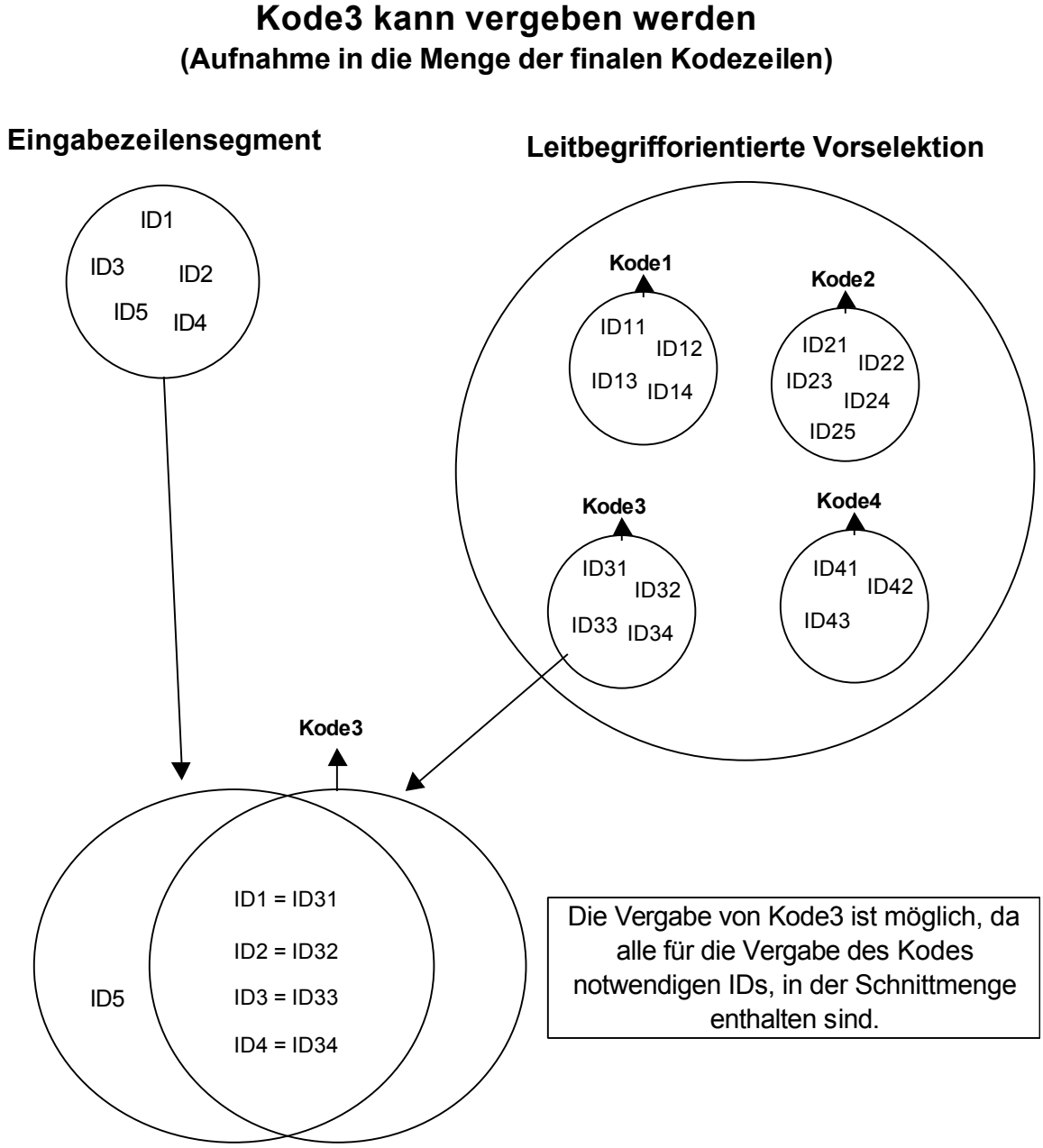

Abbildung 11: Beispiel: Kodevergabe möglich

Falls die vorstehend beschriebene Teilmengen-Beziehung für kein Element der Vorselektion erfüllt ist, mißlingt die automatische Kodierung auf Basis der erfolgten Vorverarbeitung. Die Menge der finalen Kodezeilen ist leer. Es folgt die Generierung von Fehlerhinweisen, die in einem nachfolgenden Abschnitt ausführlich dargestellt wird.

Geht man an dieser Stelle von einer erfolgreichen Ermittlung mindestens einer finalen Kodezeile aus, so stellt sich das erreichte Zwischenergebnis wie folgt dar:
Aus der Menge der Kodezeilen der Vorselektion sind nunmehr diejenigen als finale Kodezeilen ausgewählt, die nur solche IDs enthalten, die in dem zu untersuchenden Eingabezeilensegment vorhanden sind. Alle diese Kodezeilen repräsentieren somit Diagnosekodes, die für das gerade bearbeitete Segment auf Basis der vorliegenden Zwischenergebnisse zugewiesen werden könnten.
In einem letzten Schritt wird aus der jetzt vorliegenden Menge der möglichen Kodes der nach heuristischen Kriterien am besten geeignete herausgefiltert.[100] Die hierbei nacheinander angewandten Kriterien lassen sich wie folgt beschreiben:

Stelligkeit des Kodes

Man geht in diesem Zusammenhang davon aus, daß mit jeder einem Kode hinzugefügten Stelle eine Spezialisierung der repräsentierten Information erfolgt. Geht man weiterhin davon aus, daß man bei der Kodierung medizinischer Diagnosen stets an einer möglichst präzisen Information interessiert ist, so läßt sich leicht ableiten, daß in dem vorliegenden Bearbeitungsschritt der Kode mit den meisten Stellen bevorzugt berücksichtigt wird.
Erfüllt genau ein Kode das vorgenannte Kriterium, so ist das Verfahren an dieser Stelle beendet. Für das vorliegende Eingabezeilensegment wurde auf Basis der erfolgten Vorverarbeitung ein geeigneter Diagnosekode gefunden.
Erfüllen mehrere verschiedene Kodes das vorgenannte Kriterium, so fährt man im Rahmen der Optimierung mit der Evaluation des folgenden Kriteriums fort:

Anzahl berücksichtigter Modifikatoren

Bei diesem heuristischen Optimierungskriterium geht man davon aus, daß durch bei der Kodeermittlung berücksichtigte Modifikatoren Zusatzinformationen in die Kodierung einfließen, die die Präzision und Qualität der Kodeermittlung steigern. Man schließt somit von der Anzahl der berücksichtigten Modifikatoren auf die hieraus resultierende Informationsqualität.

[100] Man beachte an dieser Stelle, daß ein Diagnosekode durchaus von mehreren verschiedenen Kodezeilen repräsentiert werden kann. Die folgenden Schritte zur Optimierung beziehen sich aus diesem Grunde auf die Optimierung der Auswahl möglicher Kodes und nicht der Auswahl möglicher Kodezeilen.

Aus einer Menge auf Basis der bisherigen Bearbeitung und Optimierung geeigneter Kodes mit gleicher Stelligkeit wählt man denjenigen Kode aus, der in seiner ID-Menge die größte Anzahl von Modifikatoren enthält.
Läßt sich an dieser Stelle genau ein Kode ermitteln, so ist das Verfahren hier beendet. Für das vorliegende Eingabezeilensegment wurde bei zusätzlicher Berücksichtigung eines weiteren Optimierungskriteriums ein geeigneter Diagnosekode gefunden.

Falls nach Anwendung des vorstehend beschriebenen Optimierungsansatzes immer noch mehrere verschiedene Kodezeilen mit verschiedenen Kodes zur Auswahl stehen, muß ein finales heuristisches Entscheidungskriterium Anwendung finden:

Eine Diskussion aller möglichen Ansätze würde den Rahmen der vorliegenden Arbeit sprengen. Um dennoch zu einer eindeutigen Lösung zu kommen, wird an dieser Stelle die abschließende Auswahl des ersten Kodes auf Basis einer alphabetisch-numerisch aufsteigenden Sortierung gefordert.

Somit ist, bei entsprechendem Zwischenergebnis, in allen Fällen die Ermittlung genau eines Diagnosekodes sichergestellt.

4.2.3.2 Ermittlung von Kombikodes

Die Diagnosekodierung nach ICD-10 sieht vor[101], daß in bestimmten Fällen beim gleichzeitigen Auftraten mehrerer Diagnosen die einzelnen Diagnosekodes zu einem gemeinsamen „Kombikode“ zusammengefaßt werden. Folgendes Beispiel verdeutlicht diesen Sachverhalt:

Diagnose: „Entzündung der Mitralklappe“
ICD-10-Kode: I05.9

Diagnose: „Entzündung der Aortenklappe“
ICD-10-Kode: I35.8

Diagnose: „Entzündung der Aortenklappe mit Entzündung der Mitralklappe“
ICD-10-Kode: I08.0

[101] Vgl. [ICD10 03b].

Möchte man derartige Kombikodes bei der automatischen Diagnosekodierung berücksichtigen, so ergeben sich grundsätzlich zwei Möglichkeiten:

Alternative 1

Man versucht zunächst, aus dem Text ausschließlich Einzelkodes zu ermitteln. In einem anschließenden Analyseschritt prüft man, ob eine Kombination der aufgefundenen Einzelkodes zur Vergaben eines Kombikodes führt. Diese Alternative wurde im Rahmen der Vorarbeiten zur Erstellung von XDIAG verworfen.

Realisiert wurde vielmehr folgende Strategie:

Alternative 2

Im Datenbestand des IDT sind sowohl Einzelkodes als auch Kombikodes repräsentiert. Aus diesem Grunde ist es möglich, Kombikodes auf die gleiche Art und Weise zu bearbeiten und aufzufinden wie Einzelkodes, da entsprechende Kodezeilen in der Kodemenge vorhanden sind. Die vorstehend beschriebene Ermittlung von Einzelkodes führt bei entsprechender Gestaltung der Kodemenge somit ohne jeden funktionellen Mehraufwand auch zur Bestimmung von Kombikodes. Die in diesem Zusammenhang eingeführten Kriterien für die abschließende Kodeauswahl vermeiden hierbei, daß ein oder mehrere Einzelkodes an Stelle des Kombikodes vergeben werden.

4.2.4 Die Generierung von Fehlerhinweisen

Wie bereits beschrieben, erfolgt die Generierung von Fehlerhinweisen durch den realisierten Prototypen in den Fällen, in denen die automatische Ermittlung von Kodes mißlingt. Zu diesem Zeitpunkt hat man somit bereits festgestellt, daß die ID-Menge keines Elementes der Vorselektion eine Teilmenge des gerade zu untersuchenden Eingabezeilensegmentes ist.
Zur Generierung von Fehlerhinweisen bildet man nunmehr die Differenzmengen der ID-Mengen der einzelnen Elemente der Vorselektion mit dem Eingabezeilensegment. Eine bestimmte Differenzmenge charakterisiert somit die Vorzugsbegriffe (repräsentiert durch deren IDs), die im gerade untersuchten Eingabezeilensegment fehlen, um den Kode eines bestimmten Kode-Kandidaten der Vorselektion vergeben zu können.

Durch geschickte und praxisgerechte Darstellung dieser fehlenden Vorzugsbegriffe können dem Benutzer Informationen über im Kodierungszusammenhang notwendige

Ergänzungen präsentiert werden. Dieser Ansatz repräsentiert ein „Checklisten-Konzept“, da der Benutzer notwendige Ergänzungen nicht erst selbst „aus dem freien Raum heraus“ ermitteln muß, sondern die sich aus den bereits vorhandenen Informationen ergebenden Möglichkeiten in Form einer „Checkliste“ abprüfen kann.

Mit Blick auf eine praxisgerechte Gestaltung vorstehend beschriebener Funktionalität erscheint es aber sinnvoll, nur ausgewählte Differenzmengen zur Basis der präsentierten Fehlerhinweise bzw. Ergänzungs-Checklisten zu machen, da sonst der Benutzer durch eine zu große Informationsmenge möglicherweise verwirrt wird.

Folgende heuristische Kriterien sollten, erneut basierend auf dem Leitbegriff-Konzept, bei der Auswahl der dem Benutzer anzuzeigenden Differenzmengen herangezogen werden:

- Die Differenzmenge darf keine Leitbegriffe enthalten, da Leitbegriffe Determinanten der aufzufindenden Diagnosekodes sind und beim Fehlen dieser Determinanten der gesamte vorliegende Kodierungszusammenhang ungenügend definiert ist.
- Die Differenzmenge sollte eine möglichst kleine Anzahl von Elementen enthalten.
- Die Stelligkeit des nach Ergänzung der Differenzmenge erhaltenen Kodes sollte möglichst hoch sein.

Ziel der vorstehend beschriebenen Generierung von Fehlerhinweisen für den Benutzer ist also die strukturierte Übermittlung potentiell sinnvoller Kodierungszusammenhänge, um auf diese Weise im Sinne eines „zweiten Blickes“ ergänzende Hinweise zur korrekten und vollständigen Kodeermittlung zu bekommen.

4.3 Struktur und Erstellung der Datenbasis

Nachdem in den vorstehenden Abschnitten die Funktionalität des Prototypen detailliert erläutert wurde, soll im folgenden aufgezeigt werden, auf welche Weise die zur Realisierung der beschriebenen Funktionalität erforderlichen Eingangsdaten bereitgestellt werden können. Weiterhin wird die Struktur eben dieser Daten charakterisiert, um auf diese Weise ein besseres Verständnis unterschiedlicher im Rahmen der Evaluation diskutierter Aspekte im Zusammenhang mit notwendiger Datenpflege zu ermöglichen.

Auch an dieser Stelle soll aber ausdrücklich darauf hingewiesen werden, daß die Beschreibung der notwendigen Verfahren und Strukturen auf einer abstrakten Ebene und somit abgelöst von konkreten Implementierungsaspekten erfolgt.

Die Menge der zur Realisierung der Funktionalität von XDIAG notwendigen Daten läßt sich wie folgt gliedern:

- Daten zur *Aufbereitung* der zu kodierenden Texte
- Daten zu *Kodierung* der aufbereiteten Texte

Zum besseren Verständnis der folgenden Ausführungen soll diese Gliederung zunächst wie folgt tabellarisch in Form einer Übersicht dargestellt werden:

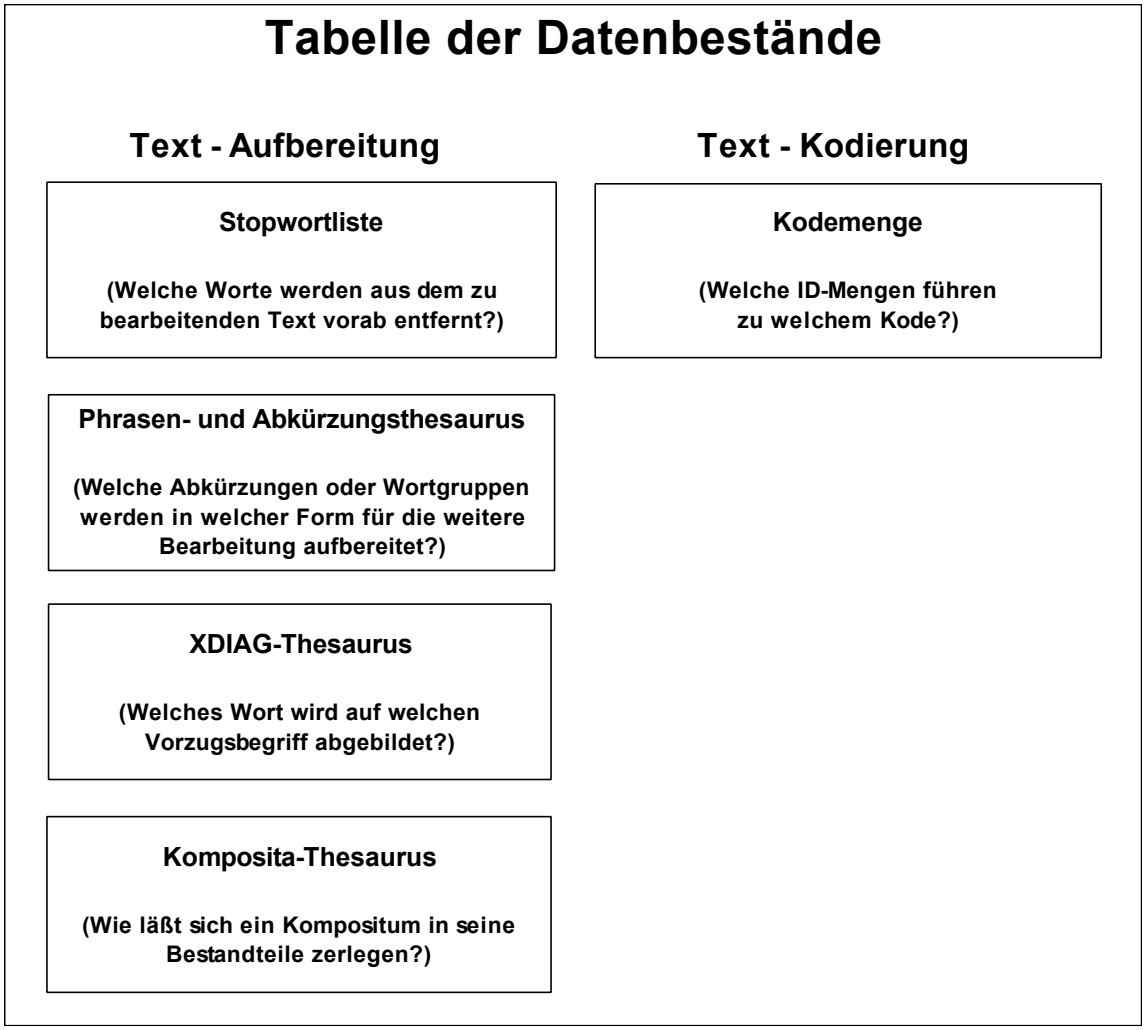

Tabelle 3: Datenbestände des XDIAG-Prototypen

4.3.1 Die Stopwortliste

Die Stopwortliste ist eine einfache Auflistung von Worten, die wie bereits beschrieben, vor allen anderen Bearbeitungsschritten aus dem zu kodierenden Text entfernt werden. Da an dieser Stelle über einen einfachen Wortvergleich hinaus keinerlei Zusatzfunktionen benötigt werden, ist die Realisierung der Stopwortliste in der Praxis besonders einfach. Eine solche Auflistung ist natürlich auch besonders einfach zu initialisieren: Durch jahrelange Erfahrung, basierend auf der Arbeit mit medizinischem Textmaterial existiert am Zentrum der Medizinischen Informatik, Frankfurt eine umfangreiche Sammlung von Stopworten. Diese Sammlung wurde für die Zwecke des vorgestellten Prototypen schließlich nur noch in Listenform überführt.

4.3.2 Der Phrasen- bzw. Abkürzungsthesaurus

Die beschriebene Auflösung von Abkürzungen bzw. die Vorverarbeitung von Phrasen erfolgt auf Basis eines speziellen Thesaurus. Dieser Thesaurus enthält nur eine Relation, die die Abkürzungen bzw. Eingangsphrasen mit den jeweils aufgelösten bzw. erweiterten Formen verbindet.[102]

An dieser Stelle ist es, wie bereits beschrieben, darüber hinaus möglich, für eine erweiterte Bearbeitung von Abkürzungen die entsprechenden Relationen mit einem Attribut zu versehen, das einen Hinweis auf den ärztlichen Bereich erlaubt, in dem die jeweilige Abkürzung relevant ist. Diese Zusatzinformation verbessert unter Umständen die Auflösung von Abkürzungen, die unterschiedliche Bedeutung haben können. Als Beispiel läßt sich an dieser Stelle nochmals die Abkürzung „HWI“ anführen: Für Internisten dürfte in der Regel „Hinterwandinfarkt“ gemeint sein; für Urologen „Harnwegsinfektion“.

Die Daten für den Phrasen- bzw. Abkürzungsthesaurus gewinnt man im wesentlichen aus zwei Quellen:

In der medizinischen Praxis verwendete Abkürzungen erhält man durch sorgfältige Analyse entsprechender Texte.

Die für die Vorverarbeitung von Phrasen besonders wichtigen „Termini technici“ erhält man durch entsprechende Analyse und Aufbereitung der Daten des IDT.

---

[102] In der Praxis stellt sich ein solcher Thesaurus mit nur einer Relation als Tabelle mit zwei Spalten dar. In der ersten Spalte stehen die Eingangsworte bzw. –wortgruppen – in der zweiten Spalten deren aufgelöste bzw. erweiterte Formen.

Weitere Phrasen, für die eine Vorverarbeitung notwendig bzw. sinnvoll sind, erhält man durch die entsprechend ausgerichtete Analyse medizinischer Texte. An dieser Stelle ist auch eine Einarbeitung des Datengutes geeigneter Lehrbücher[103] denkbar.

4.3.3 Der XDIAG-Thesaurus

Zur Realisierung von XDIAG wurde ein eigener Thesaurus entwickelt, der im folgenden als XDIAG-Thesaurus bezeichnet werden soll. Dieser XDIAG-Thesaurus bildet die Grundlage folgender vorstehend bereits beschriebener Verarbeitungsschritte:

- Abbildung von Worten auf Vorzugsbegriffe
- Auflösung bekannter Komposita
- Identifikation von Leitbegriffen

Der Thesaurus läßt sich mit Blick auf Struktur und Inhalt wie folgt beschreiben:

> Der XDIAG-Thesaurus ist, wie jeder andere Thesaurus definitionsgemäß auch, ein semantisches Netz. Für die Zwecke des in der vorliegenden Arbeit dargestellten Prototypen realisiert man die Relationen „Vorzugsbegriff“, „Kompositum“, „Leitbegriff“ sowie „POS“ = (Part of speech)[104].

Mit Hilfe dieser Relationen lassen sich die in einem vorstehenden Abschnitt der vorliegenden Arbeit beschriebenen Verarbeitungsschritte wie folgt ausführen:

Abbildung auf Vorzugsbegriffe

Soll ein Wort auf seinen Vorzugsbegriff abgebildet werden, so wird geprüft, ob für eben dieses Wort ein Thesauruseintrag vorliegt. Liegt kein solcher Eintrag vor, so ist die geforderte Abbildung nicht möglich. Das entsprechende Wort kann bei der weiteren Verarbeitung nicht berücksichtigt werden und muß einer gesonderten Behandlung zugeführt werden. Liegt ein Eintrag vor, so wird auf Basis eben dieses Eintrags die Relation „Vorzugsbegriff“ verfolgt und auf diese Weise der Vorzugsbegriff ermittelt und für die weitere Verarbeitung bereitgestellt.

---

[103] Man denke beispielhaft an die Daten eines anatomischen Terminologie-Lehrbuches.

[104] Für die Zwecke der vorliegenden Arbeit wird der zunächst sehr allgemeine Ansatz der POS auf folgende Kategorien reduziert: Diagnose; Lokalisation; Modifikator.

Es hat sich in der Praxis als vorteilhaft erwiesen, bei der Arbeit mit einem Thesaurus ausschließlich mit eindeutigen numerischen Identifikatoren von Einträgen zu hantieren. Um dies in der Praxis zu realisieren, verwendet man häufig eine „Wortliste“, in der zu jedem Thesauruseintrag dessen eindeutiger Identifikator aufgeführt ist

Als Ergebnis der Suche nach einem Vorzugsbegriff erhält man somit einen eindeutigen numerischen Identifikator. Man erkennt an dieser Stelle leicht, daß auf diese Weise eben die bereits im Zusammenhang mit dem Ablauf der Kodierung vorstehend erwähnten „IDs“ erzeugt werden.

In konzeptioneller Hinsicht deckt sich die Einführung einer „Wortliste“ mit dem von *Schalck* im Rahmen seiner Vorschläge geforderten Einzelwortkatalog.

Auflösung bekannter Komposita

Nach der Abbildung der Eingabezeile auf Vorzugsbegriffe wird in einem nächsten Schritt für jede ID der Eingabezeile geprüft, ob diese ID Teil einer Relation ist, die ein Kompositum definiert. Liegt eine solche Relation vor, wird das durch die ID repräsentierte Kompositum wie folgt aufgelöst:
Die Relation „Kompositum“ wird, ausgehend von der gerade zu untersuchenden ID ausgewertet. Die hierbei ermittelten Komponenten des Kompositums ersetzen für die weitere Verarbeitung die ursprüngliche ID. Die Zerlegung eine Kompositums in die Menge seiner Bestandteile ist somit realisiert.
An dieser Stelle soll nochmals ausdrücklich darauf hingewiesen werden, daß die Auflösung der Komposita erst auf der Ebene der Vorzugsbegriffe erfolgt, um Redundanzen bei der Datenhaltung zu vermeiden.

Identifikation von Leitbegriffen

Die Identifikation von Leitbegriffen erfolgt ebenfalls auf der Ebene der Vorzugsbegriffe. Für jede zu untersuchende ID wird geprüft, ob eine Relation Leitbegriff existiert. Wenn eine derartige Relation existiert, ist die zu untersuchende ID als Leitbegriff identifiziert.

Nachdem nunmehr der XDIAG-Thesaurus kurz charakterisiert wurde, sollen abschließend die Quellen der aktuell eingesetzten Thesaurusdaten aufgezeigt werden:

Wichtigste Grundlage für die Daten des XDIAG-Thesaurus ist der XMED-Thesaurus, der am Zentrum der Medizinischen Informatik, Frankfurt im Rahmen einer Dissertation von *Luz*[105] erstellt wurde.
Der XMED-Thesaurus basiert inhaltlich auf dem repräsentierten Wissen des AGK-Thesaurus. Der AGK-Thesaurus wurde von *Röttger*[106] ursprünglich zur medizinischen Klartextanalyse entwickelt und ist pathoanatomisch orientiert[107].

Wortliste und Relationen „Vorzugsbegriff“ und „POS“

Bei Betrachtung des AGK- bzw. XMED-Thesaurus erkennt man, daß diese in ihrer ursprünglichen Form zahlreiche Relationen beinhalten, die für die Zwecke des realisierten Prototypen nicht notwendig sind. Aus dem XMED-Thesaurus werden aus diesem Grunde als Basis des XDIAG-Thesaurus nur die Wortliste sowie die Relationen „Vorzugsbegriff“ und „POS“ unmittelbar automatisch abgeleitet.

Die Relation „Kompositum“

Die Relation „Kompositum“ ist im XMED-Thesaurus nicht vorhanden und muß somit vollständig nachgepflegt werden. Die praktische Herausforderung an dieser Stelle besteht darin, zunächst alle Vorzugsbegriffe aufzufinden, die als Kompositum behandelt werden müssen und anschließend für eben diese Vorzugsbegriffe eine mit Blick auf die Aufgabenstellung vollständige und korrekte Zerlegung zu ermitteln. Hieraus ist schließlich die Relation „Kompositum“ zu generieren.
Die an dieser Stelle benötigten Daten sind im Rahmen des realisierten Prototypen das Ergebnis einer Analyse der Daten des IDT. Auf Grund der bereits beschriebenen SGML-Auszeichnung des IDT lassen sich durch Vergleich der Software- mit der Buchversion des IDT diejenigen Einträge ermitteln, in denen Komposita vorliegen. In einem zweiten Schritt lassen sich anschließend die zur halbautomatischen Erzeugung der entsprechenden Relationen benötigten Informationen ableiten.
Die Verwendung des IDT zur Erstellung der Datenbasis zur Kompositaauflösung ist, wie bereits erwähnt, besonders deshalb sinnvoll, da der IDT eben die medizinische Sprache aus der Praxis abbildet und somit auch mit Blick auf Komposita von einer hohen Vollständigkeit und Relevanz ausgegangen werden kann.

---

[105] Vgl. [LUZ 97].
[106] Vgl. [RÖTTGER 73a] sowie [RÖTTGER 82].
[107] Vgl. [SCHALCK 73].

Die Relation „Leitbegriff“

Die Relation „Leitbegriff“ ist im XMED-Thesaurus nicht vorhanden und muß somit ebenfalls vollständig nachgepflegt werden.

In der Praxis besteht die Herausforderung an dieser Stelle darin, alle die Vorzugsbegriffe aufzufinden, die bei der Formulierung medizinischer Diagnosen im Rahmen entsprechender Texte als Leitbegriffe im Sinne des von *Schalck* eingeführten Ansatzes betrachtet werden können. Es liegt mit Blick auf Praxisnähe und Aktualität der Leitbegriffbestimmung unmittelbar auf der Hand, daß auch an dieser Stelle die Daten des IDT in den Analyseprozeß einbezogen werden müssen.

Bei der praktischen Realisierung der Leitbegriffbestimmung auf Basis der IDT-Daten stellt man fest, daß hierzu Bearbeitungsschritte notwendig sind, die auch im Rahmen der Generierung der Kodemenge anfallen. Konkret ist die Erstellung der Relation „Leitbegriff“ ein „Nebenprodukt“ der Generierung der Kodemenge und wird aus diesem Grunde im nachfolgenden Abschnitt der vorliegenden Arbeit beschrieben.

4.3.4 Die Kodemenge

Ziel des nachfolgenden Abschnittes ist es, aufzuzeigen, auf welche Weise die für die vorstehend beschriebenen Verarbeitungsschritte von XDIAG benötigte Kodemenge realisiert werden kann. Hierbei werden sowohl Datenherkunft als auch Datenrepräsentation aufgezeigt.

Hält man sich an dieser Stelle nochmals das bereits dargestellte Konzept *Schalcks* vor Augen, so erkennt man, daß die Kodemenge den von *Schalck* geforderten Schlüsselkatalog repräsentiert.

In einem vorstehenden Abschnitt der vorliegenden Arbeit wurde bereits die Verwendung der Daten des IDT als Stammdatenbasis motiviert. Es liegt somit auf der Hand, daß die Kodemenge aus den Daten des IDT extrahiert werden muß. Hierbei erscheint es sinnvoll, die Extraktion der Daten in einer Form zu gestalten, die es ermöglicht, bei Änderungen des IDT die Datenbasis mit vernünftigem Aufwand auf den aktuellen Stand zu bringen und somit das auf Basis des Prototypen realisierte System zur automatischen Diagnosekodierung stets mit einer aktuellen und damit "lebenden" Datenbasis zu versorgen.

Die im Rahmen von XDIAG realisierte Aufbereitung der IDT-Daten läßt sich wie folgt beschreiben:

Schritt 1:
Die Softwareversion des IDT wird zeilenweise für die weiteren Bearbeitungsschritte bereitgestellt. Wichtig hierbei ist, daß in einer jeden Zeile sowohl der jeweilige IDT-Diagnosetext als auch der entsprechende Diagnosekode eindeutig identifizierbar sind.

Man findet im IDT beispielsweise folgende Zeile:

J02.9; Akute Halsentzündung

Schritt 2:
Die Weiterverarbeitung erfolgt in mehreren Schritten, wobei diese Schritte mit den Einzelschritten im Rahmen der Vorverarbeitung der vom Prototypen zu analysierenden medizinischen Freitexte identisch sind, wobei allerdings mit Blick auf eine höhere Zuverlässigkeit und Präzision auf die „unscharfe Suche“ (Schreibfehlerkorrektur) verzichtet werden muß. Folgende Schritte sind somit auszuführen:

- Entfernen der Stopworte
- Abkürzungs- bzw. Phrasen-Vorverarbeitung
- Abbildung auf Vorzugsbegriffe
- Kompositaauflösung

Als Resultat der vorstehend beschriebenen Aufbereitungsfunktionen ergeben sich nunmehr Zeilen mit folgender Struktur:

ICD-10-Kode; ID1, ID2, ID3, ...IDn

Das in Schritt 1 eingeführte Beispiel stellt sich somit wie folgt bearbeitet dar:

J02.9; 452, 7895, 4568[108]

[108] Zum besseren Verständnis: 452 = ID für “akut”, 7895 = ID für “Hals”, 4568 = ID für “Entzündung”.

Schritt 3:

Abschließend muß für jede Zeile der Leitbegriff gekennzeichnet werden. Hierfür müssen als bereits vorhanden vorausgesetzte Informationen des XDIAG-Thesaurus herangezogen werden. Als Ergebnis erhält man einerseits für jede Zeile den Leitbegriff und andererseits als Vereinigungsmenge der Leitbegriffe aller Zeilen eine Menge von IDs, aus denen sich die bereits im vorstehenden Abschnitt beschrieben Relation „Leitbegriff“ (als Vereinigungsmenge aller Leitbegriffe) automatisch ableiten läßt.

Zur Bestimmung des Leitbegriffs einer IDT-Zeile nutzt man folgende heuristische Regelmenge, die aus dem einführend beschriebenen Leitbegriff-Konzept von *Schalck* abgeleitet ist:

- Enthält eine IDT-Zeile eine ID, die eine *Diagnose* beschreibt, so ist diese ID bzw. der korrespondierende Vorzugsbegriff der Leitbegriff dieser IDT-Zeile. Enthält eine IDT-Zeile mehrere Diagnosen, so wählt man die erste aufgefundene.
- Enthält eine IDT-Zeile keine ID, die eine Diagnose beschreibt, so ist zu prüfen, ob in der entsprechenden Zeile eine ID enthalten ist, die eine *Lokalisation* beschreibt. Wird eine Lokalisation gefunden, so ist diese ID bzw. der korrespondierende Vorzugsbegriff der Leitbegriff dieser IDT-Zeile. Bei mehreren Lokalisationen wählt man die erste aufgefundene.[109]

Natürlich kann an dieser Stelle theoretisch auch der Fall auftreten, daß in einer IDT-Zeile weder eine Diagnose noch eine Lokalisation angegeben ist. Ein solcher Fall muß im Rahmen der Leitbegriffbestimmung entsprechend protokolliert und fachgerecht nachbearbeitet werden. Im Verlaufe einer derartigen Nachbearbeitung wäre beispielsweise zu prüfen, ob durch entsprechende „Auslegung“ des Begriffes „Lokalisation“ nicht doch eine Lokalisation aufgefunden werden kann.[110]

[109] Man halte sich an dieser Stelle beispielsweise die IDT-Zeile „Akutes Abdomen“ vor Augen.

[110] Man denke in diesem Zusammenhang an folgende IDT-Zeile: „Abnormer Mineral-Blutwert; R79.0“ [DIMDI 01]. In dieser Zeile kann nur dann ein Leitbegriff aufgefunden werden, wenn „Mineral-Blutwert“ als Lokalisation betrachtet wird, obwohl diese Betrachtung auf den ersten Blick möglicherweise etwas ungewöhnlich erscheint.

Zusammenfassend ergibt sich somit für die Bestimmung der Leitbegriffe der IDT-Zeilen folgender praktischer Ablauf:

- Aufbauend auf dem Ergebnis aus Schritt 3 wird für jede ID einer vorverarbeiteten IDT-Zeile mit Hilfe des XDIAG-Thesaurus die „POS“ bestimmt.
- Der erste Vorzugsbegriff im Rahmen einer linearen Suche, der eine Diagnose beschreibt, wird als Leitbegriff der aktuell bearbeiteten IDT-Zeile gekennzeichnet.
- Findet man in der zu prüfenden Zeile keinen Vorzugsbegriff, der eine Diagnose repräsentiert, so wird der Startpunkt der Suche wieder auf den Anfang der Zeile zurückgesetzt. Der erste Vorzugsbegriff im Rahmen einer neuen linearen Suche, der eine Lokalisation beschreibt, wird als Leitbegriff der aktuell bearbeiteten IDT-Zeile gekennzeichnet.
- Findet man in der zu prüfenden Zeile auch keinen Vorzugsbegriff, der eine Lokalisationen kennzeichnet, so erfolgt eine Fehlerprotokollierung.

Abschließend wird bei erfolgreicher Leitbegriff-Bestimmung der gerade ermittelte Leitbegriff der aktuellen Zeile zur Vereinigungsmenge der Leitbegriffe aller Zeilen hinzugefügt.

Als Ergebnis der vorstehend beschriebenen Arbeitsschritte liegen nunmehr folgende Datenmengen vor:

1. Die Kodemenge, die die zur jeweiligen Kodevergabe notwendigen IDs zusammen mit den entsprechenden Kodes dokumentiert und die korrespondierenden Leitbegriffe auszeichnet.
   Somit liegen die für die Kodevergabe im Rahmen der Analysefunktionen von XDIAG notwendigen Kodezeilen (als Elemente der Kodemenge) vor.
2. Die Relation „Leitbegriff“, mit deren Hilfe für jeden Vorzugsbegriff des XDIAG-Thesaurus festgestellt werden kann, ob der entsprechende Vorzugsbegriff im Rahmen des IDT als Leitbegriff auftritt.[111]

[111] Diese Formulierung gilt zunächst natürlich nur für den initialen Zustand des Prototypen. Hat man die Datenbasis im Rahmen einer regelmäßigen Datenpflege angepaßt, so charakterisiert die Relation „Leitbegriff“ alle Vorzugsbegriffe des XDIAG-Thesaurus, die im Rahmen des IDT oder im Rahmen zusätzlich gewählter Referenzdaten als Leitbegriffe aufgetreten sind bzw. auftreten.

5. Evaluation

Der in den vorstehenden Abschnitten vorgestellte Prototyp realisiert, als Ganzes betrachtet, eine leitbegrifforientierte Kodierung von in medizinischen Freitexten enthaltenen Diagnosen anhand einer vorgegebenen Diagnoseschlüsselsystematik.

Bereits bei der Vorstellung des realisierten Prototypen wurde deutlich, daß die erreichbare Kodierqualität besonders stark durch die Qualität der hinterlegten Stammdatenbasis bestimmt wird. Naturgemäß kann die Datenbasis eines Prototypen ebenfalls nur prototypischen Charakter haben. Der Vergleich mit Systemen, die auf eine jahrelang gewachsene und bereits vielfach in der Praxis validierte Datenbasis zurückgreifen, würde somit zu wertlosen Ergebnissen hinsichtlich des Leistungspotentials von XDIAG führen.

Im Rahmen eines derartigen Systemvergleichs ist darüber hinaus zu berücksichtigen, daß für die medizinische Diagnosekodierung kein echter „Gold-Standard“ existiert.[112]

Aus diesem Grunde wird in den folgenden Abschnitten eine Evaluation vorgelegt, die die bei der Erstellung von XDIAG realisierten Grundkonzepte eingehend anhand ausgewählter relevanter Kriterien evaluiert. Eine besondere Rolle spielt in diesem Zusammenhang natürlich das Grundkonzept der „Leitbegrifforientierung“.

Auf Basis der im Verlaufe der Evaluation herausgearbeiteten Ergebnisse kann abschließend auf die grundsätzliche Eignung sowie auf das Leistungspotential der angewandten Lösungsansätze geschlossen werden.

Die folgende Evaluation gliedert sich in vier Abschnitte:

1. Evaluation *allgemeiner Anforderungen.*

2. Evaluation der Berücksichtigung *formaler Charakteristika* medizinischer Texte, die im zweiten Kapitel der vorliegenden Arbeit als Entwicklungsgrundlage des Prototypen beschrieben wurden.

[112] Vgl. [HASMAN 01] sowie [FRANZ 00].

3. Evaluation der Berücksichtigung *inhaltlicher Charakteristika* medizinischer Texte, die ebenfalls im zweiten Kapitel der vorliegenden Arbeit als Entwicklungsgrundlage des Prototypen beschrieben wurden.

4. Evaluation allgemeiner *kodierungsrelevanter Aspekte.*

5.1 Allgemeine Anforderungen

Einfache Datenpflege

Die Notwendigkeit permanenter Pflege der einem automatischen Kodiersystem zugrundeliegenden Daten liegt unmittelbar auf der Hand:
Einerseits ist medizinische Fachsprache ständigen Veränderungen unterworfen – andererseits werden auch Schlüsselsysteme, wie z.B. der ICD-10, an neue Entwicklungen angepaßt und somit ergänzt bzw. modifiziert.

Eben diese Notwendigkeit permanenter Datenpflege führt in der Praxis zur Forderung nach einer einfach durchzuführenden Datenpflege. Diese Forderung soll nunmehr für das prototypisch realisierte System evaluiert werden:

Diese Evaluation der Datenpflegeaktivitäten muß insbesondere mit Blick auf einen Einsatz in der medizinischen Praxis unterschiedliche Dimensionen berücksichtigen:

- Programmierung und optische Ausgestaltung von Datenpflegewerkzeugen[113]
- Inhaltliche bzw. fachliche Komplexität der Datenpflege

Da die Evaluation von Pflegewerkzeugen den Rahmen der vorliegenden Arbeit sprengen würde, wird an dieser Stelle auf allgemeine Aspekte der Softwareergonomie verweisen.[114]
Die inhaltliche Komplexität der Datenpflege wird im vorliegenden Zusammenhang durch den Aufwand charakterisiert, der anfällt, wenn Daten ergänzt oder modifiziert werden sollen.
Alle für XDIAG notwendigen Stammdaten lassen sich in Form von Tabellen darstellen. Das im System hinterlegte „kodierrelevante Wissen“ wird somit durch eben diese Tabellen repräsentiert. Die Ergänzung oder Modifikation der Datenbasis bedeutet folglich in der Praxis eine Modifikation von Tabellen.

[113] Man könnte diesen Punkt auch als „Evaluation der Datenpflegeschnittstelle“ bezeichnen.

Man kann bei den mit Dateneingabe beauftragten Personen sicher von einer gewissen Vertrautheit im Umgang mit Tabellen ausgehen[115], so daß die gewünschten Modifikationen sich ohne komplexe inhaltliche Modellierungsvorgänge umsetzen lassen.[116]
Hieraus läßt sich schließlich ableiten, daß der realisierte Prototyp die Anforderungen an eine einfache und effiziente Datenpflege erfüllt.

Aktuelle Datenbasis

Bereits bei der Einführung des IDT als geeignete Datenbasis wurde darauf hingewiesen, daß die Daten des IDT einer permanenten Pflege und Weiterentwicklung unterliegen. Durch die Verwendung dieser Daten bei der Erstellung der Datenbasis von XDIAG ist eine hohe Aktualität und Praxisnähe der verwendeten kodierrelevanten Stammdaten sichergestellt.

Qualifizierte Fehlerhinweise / Ergänzungsvorschläge

Korrekte Diagnosekodierung ist in der Praxis eine schwierige Aufgabe. Auch der Einsatz eines automatischen Kodiersystems ist grundsätzlich noch keine Garantie für vollständige und korrekte Kodes. Unmittelbar aus der Praxis läßt sich die Forderung nach einer Unterstützung durch das Kodiersystem beim Auftreten von Problemen ableiten. Wie bereits beschrieben, bietet XDIAG Hinweise in den Fällen an, in denen die aus dem Text ermittelten Informationen zur Kodierung nicht ausreichen. Die dem Benutzer angebotenen Fehlerhinweise und Ergänzungsvorschläge berücksichtigen hierbei diejenigen Informationen, die bereits aus dem Text extrahiert werden konnten und führen auf diese Weise zu einer qualifizierten Unterstützung des Benutzers:
Dem Benutzer werden nur diejenigen Ergänzungen angeboten, die die bereits vorhandenen Informationen *sinnvoll* erweitern.[117]

---

[114] Vgl. hierzu beispielsweise [SHNEIDERMAN 02].
[115] Gerade die große Verbreitung und intensive Nutzung von Standard-Software im Bereich „Tabellenkalkulation“ verstärken diese Vertrautheit.
[116] An dieser Stelle sei beispielsweise an die Darstellung des kodierrelevanten Wissens in Form einer komplexen Beschreibungssprache erinnert: Je komplexer die Beschreibungssprache, desto länger der Weg von der geplanten Änderung bis zur konkreten Manifestation eben dieser Änderung in der Datenbasis.
[117] Vgl. hierzu insbesondere das von *Giere* [GIERE 83] vorgestellte „FESCH“-System.

Ein Beispiel für einen solchen Ergänzungsvorschlag zeigt folgende Abbildung (Screenshot des Evaluations-Prototypen):

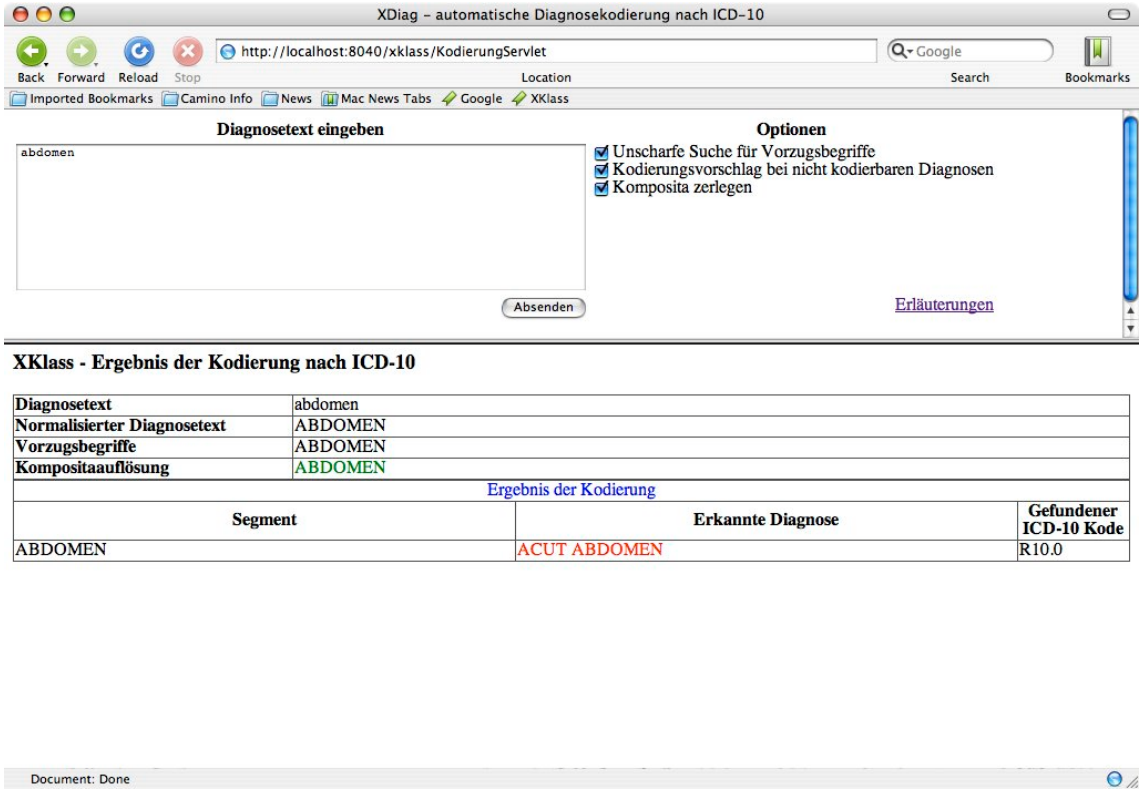

Abbildung 12: Beispiel: Ausgabe eines Ergänzungsvorschlages

5.2 Berücksichtigung formaler Charakteristika medizinischer Texte

Weitverbreitete Benutzung lateinischer und griechischer Wörter und Wortstämme

Durch die Verwendung des IDT zur Erstellung der Stammdatenbasis des Prototypen ist sichergestellt, daß in der medizinischen Praxis relevante lateinische und griechische Wörter bzw. Wortstämme im Rahmen des automatischen Kodierungsvorgangs korrekt abgearbeitet werden. In der Praxis neu gebildete bzw. neu abgeleitete kodierrelevante Wörter finden im Rahmen der routinemäßigen Datenpflege nach kurzer Zeit Eingang in den Wortschatz des IDT und können auf diese Weise auch bei der Stammdatenpflege des vorliegenden Prototypen berücksichtigt werden.

Häufige Bildung und Verwendung von Komposita

Wie bereits einleitend im Zusammenhang mit der Verarbeitung von Komposita erwähnt, besteht in diesem Zusammenhang stets die Gefahr, daß die Stammdatenbasis entweder unvollständig ist oder erhebliche Redundanzen enthält. Durch die beschriebene Vorverarbeitung der Komposita zusammen mit der anschließenden Abbildung auf Vorzugsbegriffe wird eine kompakte Datenbasis mit großem Kodierungspotential ermöglicht. Die einfache Wartung[118] der Kompositaauflösung ermöglicht darüber hinaus eine schnelle Reaktion auch auf abteilungs- bzw. personenspezifische individuelle Begriffsbildungen.

Die Abarbeitung von Komposita im Rahmen des Prototypen demonstriert folgende Abbildung (Screenshot des Evaluations-Prototypen):

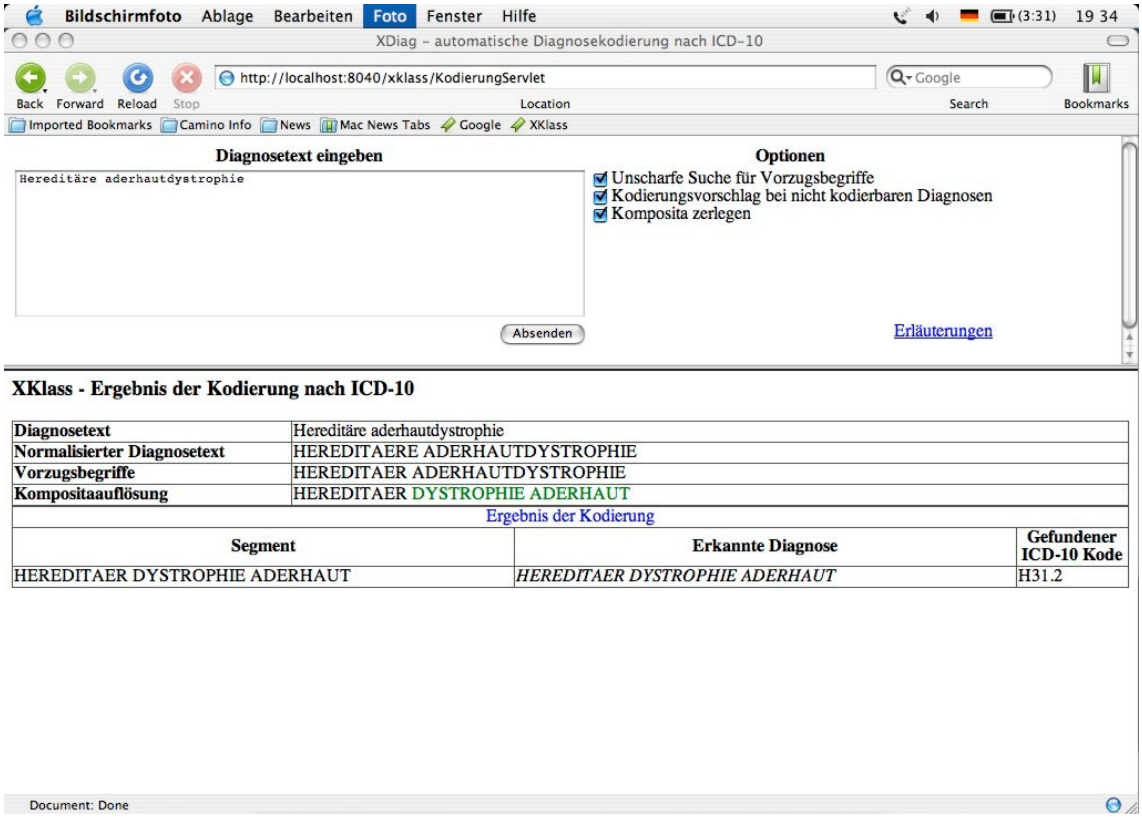

Abbildung 13: Beispiel: Abarbeitung eines Kompositums

[118] Die Wartung der Kompositaauflösung besteht nur aus der Anpassung einer Tabelle mit zwei Spalten: Eingangswort -> Zielworte.

Häufiges Auftreten von Abkürzungen

Abkürzungen stellen, wie bereits beschrieben, ein erhebliches Problem im Rahmen der Verarbeitung medizinischer Texte dar. Im Rahmen des vorliegenden Prototyps werden Abkürzungen in einer frühen Phase der Verarbeitung kontextorientiert aufgelöst und somit möglichst sinnvoll erschlossen. Dieser einfache Vorverarbeitungsschritt sorgt dafür, daß im Rahmen der weiteren Verarbeitung das Konzept der Einzelwortorientierung konsistent erhalten bleibt.

Das System ist weiterhin so offen gestaltet, daß doppeldeutige Abkürzungen fachspezifisch ausgewertet werden können. An dieser Stelle wird besonders deutlich, daß in der Vorverarbeitungsphase auch erweiterte kontextuelle Informationen einfließen können. Der eigentliche Kodiervorgang bleibt aber auch in solchen Fällen stets einzelwortorientiert.

Die Vorschriften zur Auflösung der einzelnen Abkürzungen werden ganz explizit und leicht nachvollziehbar[119] formuliert. Durch den Verzicht auf komplizierte Ausformulierungen entsprechender Regeln ist auch an dieser Stelle eine schnelle individuelle Anpassung der Daten möglich.

Vielfalt orthographischer Bezeichnungen

Die Vielfalt orthographischer Bezeichnungen im Rahmen der medizinischen Fachsprache birgt stets die Gefahr, daß die verwendete Stammdatenbasis entweder unvollständig ist oder erhebliche Redundanzen enthält[120]. Bei der Realisierung von XDIAG wird dieses Problem durch die Abbildung auf Vorzugsbegriffe gelöst. Der Einsatz des IDT zur Erstellung der Stammdatenbasis sorgt hierbei dafür, daß alle in der Praxis normalerweise verwendeten Bezeichnungen Berücksichtigung finden.

Das Konzept der Vorzugsbegriffe stellt somit im Rahmen des realisierten Prototypen sicher, daß selbst bei unterschiedlichster Begrifflichkeit automatische Kodierung anhand einer überschaubaren schlanken Stammdatenbasis möglich ist.

---

[119] Ebenfalls in Form einer zweispaltigen Tabelle.

[120] Siehe zu diesem grundsätzlichen Problem insbesondere auch die Evaluation der Kompositabehandlung.

5.3 Berücksichtigung inhaltlicher Charakteristika medizinischer Texte

Nominalsprachlicher Charakter

Der nominalsprachliche Charakter medizinischer Texte spiegelt sich in der klaren konzeptionellen Ausrichtung der Datenbasis des Prototypen wieder:
Verben finden in der kodierrelevanten Datenbasis praktisch keine Berücksichtigung. Dieses Vorgehen sichert eine schlanke Datenbasis und somit kürzere Programmlaufzeiten sowie eine leichte Wartbarkeit.

Ambinguität

Liegt bei einem Wort Ambinguität vor, so leuchtet unmittelbar ein, daß eine *einzelwortbasierte* Abbildung nur auf *einen* Vorzugsbegriff erfolgen kann, bei dem somit ebenfalls Ambinguität vorliegt. Die Auflösung eben dieser Ambinguität erfolgt im Rahmen des realisierten Prototypen erst im Rahmen der Kodeermittlung durch Berücksichtigung der im jeweiligen Kontext vorhandenen Modifikatoren. Es liegt nahe, dieses Verfahren als „indirekte Disambiguierung“ zu bezeichnen.

Folgendes Beispiel soll das Vorgehen und das somit realisierte Konzept aufzeigen:

Eingangstext:	Bruch der Leiste rechts
Abbildung auf Vorzugsbegriffe:	Bruch; Leiste; rechts

Der anschließende Vergleich mit den zu verschiedenen Kodes vorhandenen Wortmengen[121] (Abgleich mit den Kode-Kandidaten) zeigt nunmehr, daß „Bruch“ zusammen mit „Leiste“ nur dann kodiert werden kann, wenn „Bruch“ als „Hernie“ indirekt disambinguiert wird.

Der Vorteil des vorstehend aufgezeigten Verfahrens liegt erneut in der einfachen Wartbarkeit. Die Repräsentation komplexer Disambiguierungsalgorithmen entfällt. Die Festlegung, welcher begriffliche Kontext zu welcher Interpretation eines Homonyms führt, erfolgt automatisch im Rahmen der Erstellung der Kodemenge durch Definition der Wortmengen, die zu entsprechenden Kodes führen.

[121] Es wäre an dieser Stelle korrekter, von „ID-Mengen“ zu sprechen; aus Gründen der Verständlichkeit und Nachvollziehbarkeit soll hier aber erneut von der Darstellung der Wörter als IDs abstrahiert werden.

Durch die begriffliche Breite des IDT kann bereits bei Berücksichtigung des IDT als Basis der Stammdaten von einer breiten Abdeckung medizinisch relevanter Homonyme ausgegangen werden.

Varianten in der Schreibung

Varianten in der Schreibung stellen für die automatische Verarbeitung von Freitexten stets eine erhebliche Herausforderung dar. Diese Herausforderung wird noch durch die Tatsache verstärkt, daß derartige Varianten entweder gewollt oder ungewollt als Rechtschreibfehler entstehen. Wollte man alle mögliche orthographische Varianten vorsehen sowie alle möglichen Fehler antizipieren, würde dies die Datenbasis ins Unermeßliche verbreitern. Im Rahmen von XDIAG erfolgt die Abdeckung von gewollten oder ungewollten Varianten in der Schreibung durch den Einsatz der bereits beschriebenen Ähnlichkeitssuche. Auf diese Weise wird sichergestellt, daß auch Worte mit geringen orthographische Abweichungen auf ein und denselben Vorzugsbegriff abgebildet werden. Die hieraus resultierende Rechtschreibfehlertoleranz führt in der Praxis zu einer erheblichen Steigerung des Kodierpotentials ohne aufwendige Erweiterung der Datenbasis.

An dieser Stelle muß natürlich darauf hingewiesen werden, daß die Ähnlichkeitssuche stets auch mit einem gewissen Risiko behaftet ist: Ist ein Wort extrem falsch geschrieben, so kann möglicherweise die Abbildung auf einen falschen Vorzugsbegriff erfolgen, wenn durch eine unglückliche Struktur des Rechtschreibfehlers das falsch geschriebene Wort im Sinne der angewandten Ähnlichkeitssuche einem Wort aus einem völlig anderen Zusammenhang ähnlich ist. Die Feineinstellung der Ähnlichkeitssuche stellt aus diesem Grunde stets eine Gratwanderung dar: Ist die Ähnlichkeitssuche zu restriktiv eingestellt, werden bestimmte falsch geschriebene Wörter nicht zugeordnet, ist sie zu offen eingestellt, werden Wörter möglicherweise falsch zugeordnet.

Lexikalisch bedingte Synonyme

Auch durch lexikalisch bedingte Synonyme besteht grundsätzlich die Gefahr, daß die Stammdatenbasis durch redundante Informationen unnötig aufgebläht wird. Durch die Abbildung auf Vorzugsbegriffe sowohl bei der Erstellung der Stammdaten als auch bei der Bearbeitung der zu kodierenden Texte wird dieses Aufblähen vermieden. Die Abbildung auf Vorzugsbegriffe führt mit Blick auf lexikalisch bedingte Synonyme erneut zu einer schlanken und leicht wartbaren Datenbasis.

Synonymität auf Phrasenebene

Die in medizinischen Texten häufig anzutreffende Synonymität auf Phrasenebene wird im Rahmen von XDIAG abermals durch Verwendung des IDT als Stammdatenbasis bewältigt.

Bei der Erstellung der Stammdatenbasis stellt sich mit Blick auf einen möglichst kompletten Datenbestand stets die Frage, auf welche Art und Weise ein Mediziner in der Praxis einen bestimmten Sachverhalt ausdrücken könnte. Besonders wichtig an dieser Stelle ist, daß der Fokus der synonymen Struktur eine Phrase und kein einzelnes Wort ist. Gerade diese Ausrichtung auf Phrasen mit synonymer Bedeutung macht es in der Praxis besonders schwer, alle Möglichkeiten im Sinne einer optimalen Vollständigkeit zu erfassen[122]. Die Eigenschaften des IDT ermöglichen aber auch an dieser Stelle eine optimale Erfassung und Repräsentation des praktischen medizinischen Sprachgebrauchs als besondere Teilmenge der theoretisch möglichen Ausdrucksvarianten.

Syntaktisch, semantisch und pragmatisch bedingte Paraphrasen

Für syntaktisch, semantisch und pragmatisch bedingte Paraphrasen gelten ebenfalls die vorstehenden Ausführungen.

Es sei an dieser Stelle nochmals darauf hingewiesen, daß die Synonymität auf Phrasenebene sowie syntaktische, semantische und pragmatisch bedingte Paraphrasen die zentralen Argumente für die Verwendung des IDT als Stammdatenbasis darstellen.

5.4 Allgemeine kodierungsrelevante Aspekte

Kombikodes

Die Möglichkeit, Kombikodes im Rahmen des erstellten Prototypen zu ermitteln, ist abermals das Resultat der Einbindung der IDT-Daten. Der IDT enthält beispielsweise in der Version 4.0 zahlreiche Einträge, die Kombikodes repräsentieren. Als Beispiel sei an dieser Stelle nochmals auf folgenden Eintrag der Softwareversion zurückgegriffen:

Entzündung von Aortenklappe und Mitralklappe; I08.0

[122] Man halte sich an dieser Stelle die kombinatorisch „explosiv“ hohe Anzahl an Möglichkeiten vor Augen!

Neben den Kodes für die enthaltenen Einzeldiagnosen beinhaltet der IDT also, wie bereits beschrieben, den Kode für die Kombination beider Diagnosen sowie eine mögliche Formulierung eben dieser Kombination in der medizinischen Praxis. Derartige Formulierungen lassen sich leitbegrifforientiert auswerten und sind somit einer automatischen Analyse im Rahmen des vorgestellten Prototypen zugänglich.

Man erkennt an dieser Stelle besonders deutlich den Vorteil des entwickelten Verfahrens zur Segmentierung von Diagnosesätzen: Ein Ansatz, der beispielsweise „und“ als starre trennende Struktur interpretiert, kann derartige Formulierungen nicht sinnvoll abarbeiten. Die als Modul des Prototypen realisierte flexible Segmentierung läßt die Grenzen einer zu kodierenden Wortmenge offen, solange korrespondierende Einträge in der Stammdatenbasis aufgefunden werden können.

Weiterhin zeigt sich hier der Vorteil der mengenorientierten und somit reihenfolge-unabhängigen Analyse:
In Kombikodes fließen naturgemäß mehrere Diagnose-Komponenten ein. Da die Reihenfolge dieser Komponenten in der Regel für den Kombikode ohne Bedeutung ist, ermöglicht die reihenfolgeunabhängige Auswertung der Diagnosetexte eine erhebliche Flexibilität mit Verbesserung der Kodeermittlung ohne Berücksichtigung komplexer Auswertungsregeln und ohne unnötiges Aufblähen der Datenbasis[123].

Kreuz-Stern-Kodes

Für die Ermittlung von Kreuz-Stern-Kodes[124] gelten vom Prinzip her die Ausführungen des vorstehenden Abschnitts. Der IDT beinhaltet auch Formulierungen aus der ärztlichen Alltagssprache, die zu Kreuz-Stern-Kodes führen. Somit können entsprechend formulierte Diagnosetexte, wie vorstehend beschrieben, erfolgreich leitbegrifforientiert ausgewertet werden.

Komplexe Satzdarstellung

Bei der Analyse komplexer Satzdarstellungen erreicht der realisierte einzelwort-orientierte Ansatz seine Leistungsgrenze. Im Rahmen der Vorverarbeitung können nur solche Einzelworte zusammengeführt werden, deren Zusammenhang zu diesem Zeitpunkt aus deren Stellung zueinander eindeutig ermittelbar ist (z.B. „Termini Technici“:

[123] Man denke in diesem Zusammenhang insbesondere nochmals an eine leichtere Wartbarkeit der Stammdatenbasis.
[124] Vgl. [ICD10 03b].

„Arteria spinalis“). Durch diese Zusammenführung werden während des weiteren Bearbeitungsablaufs Fehler vermieden. Ab einer bestimmten Komplexität des Satzbaus ist aber der Zusammenhang von Wörtern nicht einfach durch mengenorientierte Betrachtung des unmittelbaren Umfeldes ermittelbar. An dieser Stelle soll beispielhaft auf die häufig schwer bestimmbare Reichweite von Negationen hingewiesen werden. Folgender Satz soll dies verdeutlichen:

> **„Der Verdacht auf eine Beinfraktur kann bei Berücksichtigung der vorhandenen Armfraktur nicht bestätigt werden.“**

Derartige verschachtelte Satzstrukturen können mit dem realisierten Prototypen nicht erfolgreich kodiert werden. Man kann allerdings davon ausgehen, daß solche Sätze auch mit erheblich komplexeren automatischen Analyseverfahren selbst bei Verwendung einer eingehenden Satzanalyse zur Zeit automatisch nicht wirklich sinnvoll kodiert werden können.[125]

[125] Vgl. [MOORE 94a].

6. Zusammenfassung und Ausblick

„Computer verändern die Medizin“

Manfred Gall[126] hat diese Behauptung im Jahre 1969 aufgestellt. Aus diesen Worten läßt sich die Existenz großer Erwartungen ableiten. Eine konkrete Bewertung dieser Erwartungen erfolgt zunächst nicht.

In Erweiterung der vorstehenden Aussage *Galls* hat *Giere*[127] festgehalten, daß durch den Einsatz von IT im medizinischen Bereich wohl erst dann größere Erleichterungen erwartet werden dürfen, wenn die Datenverarbeitung über die Grenzen einzelner Organisationseinheiten mit vereinheitlichter Systematik benutzt werden kann.

Medizinische Klassifikation ist eine wichtige Facette des Begriffes „vereinheitlichte Systematik“. Im zur Zeit vorherrschenden medizinischen Alltag kann man, abgesehen von der gesetzlich vorgeschriebenen Verpflichtung, bestimmte Diagnosen zu kodieren, davon ausgehen, daß vorhandene freitextliche medizinische Dokumentation in der Regel nicht in dem Maße inhaltlich ausgewertet wird, wie dies auf Basis einer „vereinheitlichten Systematik“ zur Verbesserung der Forschung oder sogar der unmittelbaren Patientenversorgung möglich wäre.[128] Eine wichtige Voraussetzung für einen effizienten und effektiven IT-Einsatz im medizinischen Bereich scheint somit in der aktuellen medizinischen Realität nicht oder nicht in angemessener Qualität vorzuliegen.

Der im Rahmen der vorliegenden Arbeit vorgestellte Prototyp setzt an dieser offensichtlich bestehenden Schwierigkeit an:

XDIAG realisiert ein IT-gestütztes leitbegrifforientiertes Verfahren zur automatischen Kodierung von Diagnosen auf Basis vorliegender medizinischer Freitexte. Die hierbei realisierten Ansätze und Verfahren folgen den Vorschlägen von Herrn *D. Schalck* und

---

[126] Vgl. [GALL 69].
[127] Vgl. [GIERE 75].
[128] Vgl. [HRIPCSAK 98] sowie [WILCOX 99].

sind somit geprägt von langjähriger intensiver und praxisnaher Beschäftigung mit Fragen medizinischer Freitextverarbeitung und Klassifikation.

Folgende Abbildung zeigt die Darstellung der Abarbeitung eines komplexeren medizinischen Diagnosetextes und soll auf diese Weise einen Gesamteindruck vom Leitungspotential der realisierten Ansätze und Verfahren vermitteln (Screenshot des Evaluations-Prototypen):

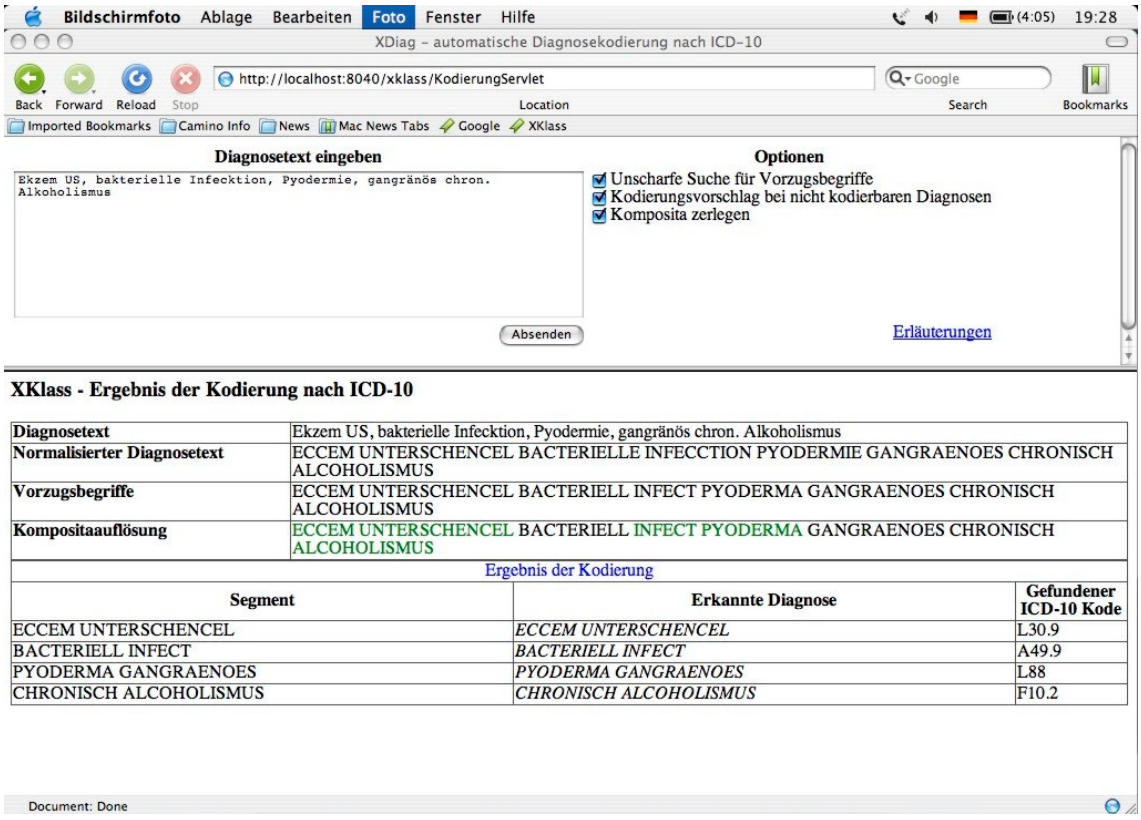

Abbildung 14: Beispiel: Kodierung eines komplexen Diagnosetextes

In Abgrenzung zu bestehenden Verfahren erfolgt eine konsequente Reduktion der Komplexität der eingesetzten Algorithmen durch einen Verzicht auf eine tiefgreifende linguistische Analyse[129] der zur Kodierung vorgelegten Texte. Dieser Verzicht und somit eine konsequente Komplexitätsreduktion sind nicht nur Teil der Ausführungen der vorliegenden Arbeit, sondern auch Teil der aktuell geführten Diskussion im Zusammenhang mit der Verarbeitung medizinischen Freitextes:

[129] An dieser Stelle sei nochmals beispielhaft die Syntaxanalyse aufgeführt.

Cimino[130] zeigt auf, daß die für komplexe medizinlinguistisch orientierte Verfahren notwendigen Daten schwer zu pflegen sind. *Friedman*[131] weist ergänzend insbesondere auf die Folgen für die Systemleistung bzw. –laufzeit hin, die die analytische Berücksichtigung eines komplexen Regelwerkes mit sich bringt.

Rector[132] hingegen führt aus, daß medizinlinguistisch orientierte Probleme mit einer begrenzten Komplexität durchaus in überschaubaren Zeiträumen und mit einer überschaubaren Datenbasis lösbar sind.

Die mit dem Ziel einer reduzierten Komplexität im Rahmen von XDIAG realisierten Lösungsansätze verleihen dem Gesamtverfahren den Charakter einer Heuristik. Hieraus lassen sich unmittelbar Stärken und Schwächen des Prototypen ableiten:

- Alle entwickelten Algorithmen nutzen die Besonderheiten der medizinischen Fachsprache, um auf diese Weise eine besonders effektive und effiziente automatische Kodierung von Diagnosen zu ermöglichen. Diese Effektivität führt zu einer im Vergleich zu syntaktisch orientierten Verfahren schlanken und leicht zu pflegenden Datenbasis. Eben diese Datenpflege kann, bedingt durch den Verzicht auf die Repräsentation komplexer Regelwerke, auch von Mitarbeitern ohne linguistische Kenntnisse effektiv und effizient durchgeführt werden.
- Es kann grundsätzlich nicht garantiert werden, daß der Prototyp für einen vorgelegten Text die optimale Diagnosekodierung findet. Die gefundene Kodierung kann aber auf Grund des leitbegrifforientierten Vorgehens als gute Näherung an den optimalen Kode betrachtet werden. Gerade im Rahmen statistischer Auswertungen erscheint dieses Vorgehen mehr als akzeptabel.
- Texte mit einer bestimmten Komplexität bzw. strukturellen Verschachtelung können nicht oder nur mit erheblich reduzierter Kodiergenauigkeit bearbeitet werden.

Auf Basis der leitbegrifforientierten Vorgehensweise realisiert XDIAG einige Funktionen, die in zahlreichen vorhandenen Systemen fehlen:

---

[130] Vgl. [CIMINO 01].
[131] Vgl. [FRIEDMAN 97].
[132] Vgl. [RECTOR 99].

- Durch die Möglichkeit, qualifizierte Fehlerhinweise mit Ergänzungsvorschlägen interaktiv zu generieren, ergibt sich für den Anwender die Option, durch wenige zusätzliche Informationen die Kodierqualität nach seinen Wünschen bzw. Bedürfnissen gezielt zu steigern.
- Durch die Möglichkeit, mehrere Diagnosen in einem Textabschnitt zu kodieren, werden die häufig durch die Punkt-zu-Punkt-Regel gesetzten funktionalen Grenzen im Sinne größerer Vollständigkeit und Genauigkeit aufgehoben.
- Mit Blick auf eine Erweiterung der Mächtigkeit des Verfahrens ist auch an eine Anbindung der Daten des „Deutschen Specialist Lexicon“[133] zu denken. Diese Anbindung dürfte, bedingt durch die einfache Datenstruktur des Prototypen, mit überschaubarem Aufwand realisierbar sein.
- Das Leistungspotential eines derartigen automatischen Verfahrens läßt sich insbesondere dann nutzen, wenn bei der Erstellung der der Analyse zugrunde liegenden Texte gewisse Grundregeln zusammen mit einem gewissem Maß an Sprachdisziplin eingehalten werden:
  Texte, bei denen Interesse an einer Weiterverwendung besteht, sollten klar und unter Vermeidung möglicher Ambiguitäten formuliert werden. Weiterhin sollte auf korrekte Orthographie sowie auf klare und korrekte Satzstrukturen geachtet werden.[134]

An der richtigen Stelle eingesetzt und mit der richtigen Motivation benutzt und gepflegt, kann der im Rahmen der vorliegenden Arbeit vorgestellte Prototyp helfen, eine synergistische Brücke zwischen praktischer Medizin, medizinischer Verwaltung und medizinischer Forschung zu schlagen.

[133] Vgl. [WESKE 02].
[134] Vgl. [MOORE 94] sowie [MOORE 94a].

7. Literaturverzeichnis

[AHO 74] Aho, A. V., Hopcroft, J. E., Ullmann, J. D.: The Design and Analysis of Computer Algorithms. Addison Wesley, Reading, Massachusetts etc., 1974.

[ALTENPOHL 74] Altenpohl, U.: Entwurf eines automatischen Diagnoseverschlüsselungssystems auf der Basis der "Sytematized Nomenclature of Pathology" (SNOP) und der Wortsegmentierung. Symposium über Klartextanalyse in der Medizin (2), 22.6.1974, SIEMENS, 45 - 52.

[BO 85] Berufsordnung für Ärzte in Hessen, Allgemeine Vorschriften. Landesärztekammer Hessen, 1985.

[BOUCHET 98] Bouchet, C., Bodenreider, O., Kohler, F.: Integration of the analytical and alphabetical ICD10 in an coding help system. Proposal of a theoretical model for the ICD representation. Medinfo, 1998, Pt 1, 176 - 179.

[BOUCHET 98a] Bouchet, C., Empereur, F., Kohler, F.: Evaluating a computerized tool for coding patient information. Proc AMIA Symp., 1998, 185 - 189.

[BURKART 90] Burkart, M.: Dokumentationssprachen. In: Buder, M.; Rehfeld, W.; Seeger, T. (Hrsg.): Grundlagen der praktischen Information und Dokumentation. Ein Handbuch zur Einführung in die fachliche Informationsarbeit. München: Saur, 3. Auflage, 1990, S. 143 - 182.

[CIMINO 96] Cimino, J. J.: Review Paper: Coding Systems in Health Care. Meth Inf in Med. 1996; 35 No. 4/5, 273 - 284.

[CIMINO 98] Cimino, J. J.: Desiderata for Controlled Medical Vocabularies in the Twenty-first century. Meth Inf in Med. 1998; 37, 394 - 403.

[CIMINO 01] Cimino, J. J.: Terminology Tools: State of the Art and Practical. Meth Inf in Med. 2001; 40 No. 4, 298 - 306.

[DEBRUIJN 97] De Bruijn, L. M., Hasman, A., Arends, J. W.: Automatic SNOMED classification - a corpus-based method. Comput Methods Programs Biomed. 1997 Sep; 54 (1-2), 115 - 122.

[DEBRUIJN 98] De Bruijn, L. M., Hasman, A., Arends, J. W.: Automatic Coding of Diagnostic Reports. Meth Inf in Med. 1998; 37, 260 - 265.

[DEIMEL 97] Deimel, D., Hesselschwerdt, H. J., Heisel, J.: Simple and valid ICD9-/10- and IKPM coding using an electronic data-assisted coding system "do it" -- experiences after one years use. Orthop Ihre Grenzgeb. 1997 Nov-Dec, 135, (6), 528 - 534.

[DICK 91] Dick, R. S., Stehen, E. B.: The Computer-Based Patient Record: An essential Technology for Health Care. Washington, DC: National Academy Press, 1991.

[DIMDI 01] Deutsches Institut für medizinische Dokumentation und Information (DIMDI): ICD-10-Diagnosenthesaurus: Sammlung von Krankheitsbegriffen im deutschen Sprachraum, verschlüsselt nach der internationalen statistischen Klassifikation der Krankheiten und verwandeter Gesundheitsprobleme – Version 4.0, Stand Januar 2001. Dt. Ärzte-Verlag, Köln, 2001.

[DOMSCHKE 93] Domschke, W.; Scholl, A.; Voß, S.: Produktionsplanung – Ablauforganisatorische Aspekte. Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1993.

[ERTEL 73] Ertel, M.: Semantische Aspekte medizinischer Information. Symposium über Klartextanalyse in der Medizin, 23.6.1973, SIEMENS, 3 - 11.

[FEIGL 73] Feigl, W.: Definitionen und Voraussetzugen der medizinischen Klartextanalyse. Symposium über Klartextanalyse in der Medizin, 22.6.1973, SIEMENS, 27 - 35.

[FEIGL 74] Feigl, W., Röttger, P., Köberl, D.: Zur Informations-Transformation bei der Übertragung von Biopsie-Diagnosen aus der Klartext-Form in den SNOP-Code. Symposium über Klartextanalyse in der Medizin (2), 22.6.1974, SIEMENS, 3 - 16.

[FRIEDMAN 97] Friedman, C.: Towards a comprehensive medical language processing system: methods and issues. Proc AMIA Annu Fall Symp. 1997; 595 -599.

[FRANZ 00] Franz, P., Zaiss, A., Schulz, S., Hahn, U., Klar, R.: Aztomated Coding of Diagnoses - Three methods compared. Proc AMIA Symp., 2000, 250 - 254.

[GALL 69] Gall, M. W.: Computer verändern die Medizin. A. W. Gentner, 1969, Stuttgart.

[GIERE 69] Giere, W., Baumann, H., et al. (1969). Der programmierte Arztbrief. Ein Weg zur klinischen Volldokumentation. Stuttgart, IBM.

[GIERE 75] Giere, W.: Projekt Datenverarbeitung in der Medizin; Einführung der Datenverarbeitung in die ärztliche Praxis – Dokumentation und Informationsverbesserung in der Praxis des niedergelassenen Arztes mittels EDV-Service (DIPAS). DVM-Bericht 3, 1975.

[GIERE 83] Giere, W.: FESCH: Fachkraftgesteuerte Erstellung von Schlüsseln (FESCH) aus Diagnosen oder Befunden im Klartext. Papier der ADD, Frankfurt/Main, 1983.

[GIERE 84] Giere, W.: Krankendaten: Dokumentation für Medizin oder Bürokratie? Medizinische Informatik und Statistik; Herausgeber: S. Koller, P.L. Reichertz und K. Überla; Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, Bd. 58, 1984.

[GIERE 86a] Giere, W.: BAIK – Befunddokumentation und Artztbriefschreibung im Krankenhaus. Media Verlag, Taunusstein, 1986.

[GIERE 86b] Giere, W.: Medizinische Dokumentation: Einführungsreferat. In: Pannike, A. (Hrsg.): Hefte zur Unfallheilkunde, Heft 181. Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1986.

[GIERE 02] Giere, W.: Prüfsteine für die digitale Patientenakte. In: Deutsches Ärzteblatt 2002; 99 [Heft 6], 344 - 346.

[GOLDFARB 90] Goldfarb, Charles F.: The SGML Handbook. Oxford University Press, 1990.

[GREGORI 95] Gregori, A.: Klartextverarbeitung als integraler Bestandteil medizinischer Dokumentation - Modell, Implementierung und Evaluation eines maschinellen Klassifikationssystems; Hochheim, Darmstadt: Epsilon-Verlag, 1995.

[HALL 86] Hall, P. A., Lemoine, N. R.: Comparison of manual data coding errors in two hospitals. Journal of Clinical Pathology, 1986; 39; 622 - 626.

[HASMAN 01] Hasman, A., De Bruijn, L. M., Arends, J. W.: Evaluation of a Method that Supports Patology Report Coding. Meth Inf in Med. 2001; 40, 293 - 297.

[HOHNLOSER 96] Hohnloser, J. H., Kadlec, P., Puerner, F.: Coding Clinical Information: Analysis of Clinicians Using Computerized Coding. Meth Inf in Med. 1996; 35 No. 2, 104 - 107.

[HOHNLOSER 96a] Hohnloser, J. H., Puerner, F., Soltanian, H.: Improving coded data entry by an electronic patient record system. Meth Inf in Med. 1996; 35 No. 2, 108 - 111.

[HRIPCSAK 98] Hripcsak, G., Kuperman, G. J., Friedman, C.: Extracting Findings from Narrative Reports: Software Transferability and Sources of Physician Disagreement. Meth Inf in Med. 1998; 37, 1 - 7.

[ICD10 03a] ICD10 - ICD-10-GM Systematisches Verzeichnis Version 2004; Krankenhausdrucke-Verlag GmbH; 1. Aufl., 2003.

[ICD10 03b] ICD10 - Allgemeine und Spezielle Kodierrichtlinien für die Verschlüsselung von Krankheiten und Prozeduren; Krankenhausdrucke-Verlag GmbH; 1. Aufl., 2003.

[ICD10 99] ICD10 – WHO – Ausgabe; Band III – Alphabetisches Verzeichnis; 10. Rev.; Vers. 1.3; Kohlhammer, 1999.

[KUCHENBECKER 95] Kuchenbecker, J., Ehrt, O., Guthoff, R.: Computer program for improved diagnostic coding in ophthalmology based on the expanded ICD 10. Klein Monatsbl Augenheilkd., 1995, Jul, 207 (1), 46 - 50.

[LAUTERBACH 00] Lauterbach, K., Lüngen, M.: Neues Entgeltsystem nach US-Muster. Deutsches Ärzteblatt, 97 (2000) A, 444 - 447.

[LOVIS 00] Lovis, C., Baud, H.: Fast exact string-matching Algorithms adapted to the characteristics of the medical language. Journal of the American Medical Informatics Association, number 7, July, 2000, 378 - 391.

[LLOYD 85] Lloyd, S. S., Rissing, J. P.: Physician and Coding Errors in Patient Records. Journal of the American Medical Informatics Association, No. 2547, 1985, 1330 - 1336.

[LLOYD 97] Lloyd, S. S., Layman, E.: The eefcts of automated encoders on coding accuracy and coding speed. Top Health Inf Manage, 1997, Feb, 17 (3), 72 - 79.

[LUZ 97] Luz, C.: XMED - vom Freitext zum Kode. Praxis der EDV-gestützten Klassifikation medizinischer Texte nach ICD und IKPM. Darmstadt, Epsilon-Verlag, 1999.

[MEHLHORN 88] Mehlhorn, K.: Datenstrukturen und effiziente Algorithmen; Bd. 1 Sortieren und Suchen. Stuttgart: Teubner, 1988.

[MOORE 89] Moore, G. W., Miller, R. E., Hutchins, G. M.: Indexing by MeSH titles of natural language pathology phrases identified on first encounter using the barrier word method. In: Computerized Natural Medical Language Processing for Knowledge Engineering. Scherrer, J. R. et al. (eds.), Elsevier Science Publishers, North Holland, 1989, 29 - 45.

[MOORE 94] Moore, G. W., Berman, J. J.: Performance analysis of manual and automated systemized nomenclature of medicine (SNOMED) coding. Am J Clin Pathol. 1994 Mar;101(3), 253 - 256.

[MOORE 94a] Moore, G. W., Berman, J. J.: Automatic SNOMED coding. Proc Annu Symp Comput Appl Med Care. 1994, 225 -229.

[NANGLE 94] Nangle B., Keane MT.: Effective retrieval in Hospital Information Systems: the use of context in answering queries to Patient Discharge Summaries. Artif Intell Med. 1994 Jun; 6(3), 207 - 227.

[NITZSCHKE 92] Nitzschke, E., Wiegand, M.: Analysis of errors in ICD 9 diagnostic classification in compliance with the Federal health care regulation. Orthop Ihre Grenzgeb. 1992 Sep-Oct, 130, (5), 371 - 377.

[RAPP 97] Rapp, R.: Text-Detector: Fault-tolerant Retrieval Made Simple. C'T Magazin für Computer-Technik; Ausgabe 4; 1997, 386ff.

[RECTOR 99] Rector, A. L.: Clinical Terminology: Why Is It so Hard? Meth Inf in Med. 1999; 38, 239 - 252.

[RÖTTGER 69] Röttger, P., Reul, H., Klein, I., Sunkel, H.: Vollautomatische Dokumentation und statistische Auswertung pathologisch-anatomischer Befundberichte. Meth Inf in Med. 1969; 8, 19 - 26.

[RÖTTGER 73] Röttger, P.: Informationstheoretische Aspekte medizinischer Routine-Befundmitteilungen. Symposium über Klartextanalyse in der Medizin, 23.6.1973, SIEMENS, 12 - 23.

[RÖTTGER 73a] Röttger, P. et. al.: Konzeption und Organisation des AGK-Thesaurus. Symposium über Klartextanalyse in der Medizin, 23.6.1973, SIEMENS, 52 - 60.

[RÖTTGER 82] Röttger, P.: Klartextverarbeitung in der Pathologie. Habilitationsschrift, Klinikum der J. W. Goethe-Universität, Frankfurt am Main, 1982.

[SCHALCK 02] Schalck, D.: Sammlung Schalck. Uniklinikum Frankfurt/Main, Frankfurt/Main, 2002.

[SCHALCK 73] Schalck, D., Arndt, F., et al.: Erfahrungen bei Anwendung des AGK-Thesaurus im Bereich der Inneren Medizin. Symposium über Klartextanalyse in der Medizin, 23.6.1973, SIEMENS, 76 - 83.

[SCHALCK 74] Schalck, D., Arndt, F., Giere W.: Die klinische Diagnose im sprachstatistischen Vergleich. Symposium über Klartextanalyse in der Medizin (2), 22.6.1974, SIEMENS, 73 - 84.

[SCHERRER 90] Scherrer J.R., Baud R.H., Hochstrasser D., Ratib O.: An integrated hospital information system in Geneva. MD Comput. 1990 Mar-Apr; 7(2), 81 - 89.

[SHNEIDERMAN 02] Shneiderman, B.: User Interface Design; Effektive Interaktion zwischen Mensch und Maschine. Leitfaden für intelligentes Schnittstellendesign; Deutsch von: Dubau, J., Willner, A.; Vmi Buch AG, Bonn-Oberkassel, 2002.

[SITTIG 94] Sittig, D. F.: Grand Challenges in Medical Informatics. J Am Med Inform Assoc. 1994; 1, 412 - 413.

[SPYNS 96] Spyns, P.: Natural language processing in medicine: an overview. Methods Inf Med. 1996 Dec; 35(4-5), 285 - 301.

[SURJAN 01] Surjan, G., Heja, G.: Indexing of medical diagnoses by word affinity method. Medinfo, 2001, 10 (Pt 1), 276 - 279.

[TALMON 02] Talmon, J. L., Hasman, A.: Medical Informatics as a Discipline at the beginning of the 21st Century. Meth Inf in Med. 2002; 41 No. 1, 4 - 7.

[VANBEMMEL 00] Van Bemmel, J. H., Musen, M. A., eds. Handbook of Medical Informatics; Springer Verlag, Heidelberg/New York, 2000.

[VANDERLEI 02] Van der Lei, J.: Closing the Loop between Clinical Practice, research, and Education: The Potential of Electronic Patient Record. Meth Inf in Med. 2002; 41 No. 1, 51 - 54.

[WEED 78] Weed, L. L.: Das problemorientierte Krankenblatt. Ins Deutsche übertragen von E. Beck. Stuttgart, Schattauer, 1978.

[WESKE 02] Weske-Heck G., Zaiss A., Zabel M., Schulz S., Giere W., Schopen M., Klar R.: The German Specialist Lexicon, Proc AMIA Symp. 2002, 884 - 888.

[WINGERT 74] Wingert, F.: Textverarbeitung in der Medizin. Symposium über Klartextanalyse in der Medizin (2), 22.6.1974, SIEMENS, 17 - 44.

[WINGERT 85] Wingert, F.: Morphologic Analysis of Compound Words. Meth Inf in Med. 1985; 24, 155 - 162.

[WINGERT 89] Wingert, F., Rothwell, D., Côté, R.: Automated Indexing into SNOMED and ICD. In: Scherrer, J. R., Côté, R., Mandil, S. H. (Hrsg.): Computerized Natural Medical Language Processing for Knowledge Representation. North-Holland, Amsterdam, 1989, 201 - 239.

[WILCOX 99] Wilcox A., Hripcsak G.: Classification algorithms applied to narrative reports. Proc AMIA Symp. 1999; 455 - 459.

[ZAISS 02] Zaiß, A. et al.: Medizinische Dokumentation, Terminologie und Linguistik; in: Lehmann, T. M., Meyer zu Bexten, E. (Hrsg.): Handbuch der Medizinische Informatik; Hanser-Verlag, München, Wien, 2002; S. 45 ff..

[ZIMMER 99] Zimmer, D. E.: So kommt der Mensch zur Sprache – 5. Auflage; Heyne Verlag, München, 1999.

[ZIPS 83] Zips, B., Giere W.: Klassifikation, befundorientierte Speicherung und Informationsgewinnung mit IATROS. Medizinische Informatik und Statistik; 50, 1983.

Zusammenfassung

In der medizinischen Praxis in Deutschland ist Klassifikation als essentieller Bestandteil der Dokumentation in vielen Bereichen durch gesetzliche Regelungen vorgeschrieben. Über diesen gesetzlich determinierten Rahmen hinaus können durch Klassifikation vergleichbar gemachte Informationen als Basis neuer wissenschaftlicher Erkenntnisse herangezogen werden und weiterhin helfen, bestehende Lehrmeinungen zu evaluieren.
Ein Blick auf die im medizinischen Umfeld vorhandene organisatorische Realisierung der Klassifikation zeigt, daß diese in der Regel von medizinisch qualifiziertem Fachpersonal neben der eigentlichen Tätigkeit durchgeführt wird. Eine Klassifikation vorhandener Dokumentationen im Sinne einer Erschließung zusätzlicher wertvoller Informationsquellen über den gesetzlichen Mindestumfang hinaus scheitert somit häufig an der organisatorisch bedingten Überlastung der eingesetzten Mitarbeiter.
Eine Unterstützung medizinischer Klassifikation in der Praxis durch den geeigneten Einsatz von Informationstechnologie (IT) erscheint somit sinnvoll und wünschenswert.

Im Rahmen der vorliegenden Arbeit wird ein entsprechender Ansatz in Form eines entwickelten Prototypen (XDIAG) vorgestellt und evaluiert.

Der entwickelte Prototyp realisiert ein IT-gestütztes leitbegrifforientiertes Verfahren zur automatischen Kodierung von Diagnosen auf Basis vorliegender medizinischer Freitexte. Die hierbei realisierten Ansätze und Verfahren folgen den Vorschlägen von Herrn *D. Schalck* und sind somit das Resultat langjähriger intensiver und praxisnaher Beschäftigung mit Fragen medizinischer Freitextverarbeitung und Klassifikation. Die besondere Vorgehensweise verleiht dem vorgestellten Prototypen den Charakter einer Heuristik.
In Abgrenzung zu zahlreichen bestehenden Verfahren erfolgt eine konsequente Reduktion der Komplexität der eingesetzten Algorithmen und Stammdaten durch einen Verzicht auf eine tiefgreifende linguistische Analyse der zur Kodierung vorgelegten Texte. Durch diesen Verzicht kann auf die Verwendung einer Grammatik und somit auf die Verwendung komplexer Stammdaten verzichtet werden. Als Stammdatenbasis werden vielmehr Datenbestände verwendet, die entweder besonders leicht zu pflegen sind oder aber ohnehin permanent im Rahmen von Langzeitprojekten gepflegt werden. An dieser Stelle spielt insbesondere der ICD10-Diagnosen-Thesaurus mit seiner umfassenden und besonders praxisorientierten Begriffsmenge eine wichtige Rolle.

In Erweiterung bestehender Verfahren bietet der vorgestellte Prototyp darüber hinaus die Möglichkeit, mehrere medizinische Diagnosen im Rahmen eines Satzes zu kodieren. Weiterhin können dem Benutzer interaktiv qualifizierte Fehlerhinweise mit dem Ziel einer verbesserten Kodierung bereitgestellt werden.

Als Ergebnis der Evaluation des realisierten Prototypen läßt sich festhalten, daß die hierbei eingesetzten Verfahren helfen können, eine synergistische Brücke zwischen praktischer Medizin, medizinischer Verwaltung und medizinischer Forschung zu schlagen, wenn sie an der richtigen Stelle und mit der richtigen Motivation eingesetzt werden.

Abstract

In many areas of medical practice in Germany the classification, an essential part of documentation, is regulated by a legal framework. Beyond this regulatory framework, classification has the ability to make comparative information possible which may be used as a basis for research and also aids the evaluation of current doctrines.

When assessing the current organisation of classification in the medical environment, it becomes apparent that this is generally performed by qualified professional staff in line with their actual job description. The classification of existing medical information using additional and useful sources of information beyond the legally required minimum, often fails due to the lack of time staff have because of heavy work load.

Subsequently, the support of medical classification in practice through the employment of appropriate Information Technology seems practical and desirable.

Due to this fact a prototype is presented to demonstrate and evaluate a system of procedures that can help to deliver the necessary kind of support.

The prototype enables an IT-supported lead-term-orientated system of procedures to automatically code diagnoses based on available medical free-texts. Here, the resulting starting points and procedures follow the suggestions made by *Mr. D. Schalck* and therefore come from of years of intensive and practically orientated research into questions of the processing of medical free-texts. This special process provides the prototype with a heuristic character.

As opposed to a vast number of existing processes the prototype enables a consequent reduction of complexity of the algorithms and master data used through the elimination of a syntactic analysis of the texts used for coding. This eliminates the need to use grammar and therefore also the need for employing complex master data. Hence, data banks are used as the basis of master data which are either easily maintained or maintained anyway within long term projects. The ICD10-Diagnoses-Thesaurus is of great importance at this point particularly due to its extensive and practically orientated number of expressions.

As an extension of existing processes the prototype offers the opportunity of coding several medical diagnoses within one sentence. The system also offers the user a means of receiving interactive and qualitative error messages in order to enable coding in a second step when coding in the first step fails due to incomplete or non-consistent

information. These error messages could also be used to improve the coding step by step.

The evaluation of the resultant prototype concludes that the processes employed have the ability to aid the building of a synergetic bridge between practised medicine, medical administration and medical research if used at the right point and with the right motivation.

Erklärung

Ich erkläre, daß ich die dem Fachbereich Medizin der Johann Wolfgang Goethe-Universität Frankfurt am Main zur Promotionsprüfung eingereichte Dissertation mit dem Titel

Automatische Diagnosekodierung mit XDIAG

Konzeption und Evaluation eines heuristischen Verfahrens zur leitbegrifforientierten automatischen Diagnosekodierung auf Basis der Daten des ICD10-Diagnosen-Thesaurus

im Institut für Dokumentation und Informationstechnologie unter Betreuung und Anleitung von Herrn Prof. em. Dr. med. Wolfgang Giere mit Unterstützung durch Dr. rer. med. Wolfgang Kirsten ohne sonstige Hilfe selbst durchgeführt und bei der Abfassung der Arbeit keine anderen als die in der Dissertation aufgeführten Hilfsmittel benutzt habe. Ich habe bisher an keiner in- oder ausländischen Universität ein Gesuch um Zulassung zur Promotion eingereicht.

Bad Camberg, 15. November 2004

Lebenslauf

Name:	Ralf Starzetz	
Wohnort:	Zur Hub 7, 65520 Bad Camberg	
Geburtsdatum:	22.03.1968	
Geburtsort:	Wiesbaden	
Eltern:	Vater:	Ewald Starzetz, Bankkaufmann
	Mutter:	Anneliese Starzetz, geb. Schmitz Verwaltungsangestellte (verstorben 1991)
Schulausbildung:	1974 bis 1978:	Grundschule Bad Camberg/Erbach
	1978 bis 1987:	Tilemannschule Limburg (altspr. Gymnasium)
Universitäre Ausbildung:	1989 bis 1996:	TU Darmstadt (Wirtschaftsinformatik)
Abschlüsse:	Sommer 1987:	Abitur
	Frühjahr 1996:	Diplom (Wirtschaftsinformatik)
Sprachkenntnisse:	Englisch (Klasse 5-13)	
	Latein (Klasse 7-13)	
Berufserfahrung:	1990 bis 1995:	Studienbegleitende Beschäftigung bei Fa. CSC Ploenzke in Kiedrich, Rheingau, im Bereich Buchhaltung und Personalwesen
	1999 bis 2003:	Wissenschaftlicher Mitarbeiter am Zentrum für medizinische Informatik am Universitätsklinikum Frankfurt/Main
	Seit 1996:	Selbständige Tätigkeit als IT-Berater in verschiedenen mittelständischen Unternehmen

Bad Camberg, 15. November 2004