Automatische, Deskriptor-basierte Unterstützung der Dokumentanalyse zur Fokussierung und Klassifizierung von Geschäftsbriefen

Stefan Dittrich, Rainer Hoch

Mai 1992

Vorwort

Das Projekt ALV (Automatisches Lesen und Verstehen) am Deutschen Forschungszentrum für Künstliche Intelligenz beschäftigt sich mit der wissensbasierten Dokumentanalyse und der Entwicklung von intelligenten Büroinformationssystemen. In diesem Rahmen werden u. a. statistische Verfahren und Methoden des klassischen Information Retrievals eingesetzt, um den ersten Grundstein für eine inhaltliche Erschließung der Dokumente zu legen.

Die vorliegende Dokumentation vergleicht existierende Verfahren des klassischen Information Retrievals und beurteilt sie hinsichtlich ihrer Einsetzbarkeit bzw. Erweiterbarkeit für eine intelligente Dokumentanalyse. Primäres Ziel der Arbeit bestand darin, rein statistische Verfahren für eine Klassifikation von Geschäftsbriefen (Anfrage, Angebot, Bestellbestätigung, Bestellung, Werbung, etc.) geschickt anzuwenden unter Ausnutzung der zuvor ermittelten Dokumentenstruktur (logisches Dokumentenmodell) und einem einfachen Modell für Nachrichtentypen (Geschäftbriefklassen). Insbesondere kommen eine umfangreiche Sammlung von Geschäftsbriefen, nachrichtentypspezifische Wortlisten, allgemeine Worthäufigkeiten des Deutschen sowie eine morphologische Komoponente zum Tragen.

Ausgangspunkt der Analyse ist eine Datenbasis von etwa 120 elektronisch vorliegenden, deutschen Geschäftsbriefen. Obwohl diese Briefdatenbasis relativ klein ist, fielen die Klassifikationsergebnisse derart ermutigend aus, so daß neue Aktivitäten in diesem Bereich sinnvoll erscheinen.

Kurzbeschreibung

Die vorliegende Diplomarbeit wurde im Rahmen des ALV-Projekts (**A**utomatisches **L**esen und **V**erstehen) am Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) erstellt. Ziel des ALV-Projektes ist die Entwicklung einer intelligenten Schnittstelle zwischen Papier und Rechner (paper-computer interface). Hierbei soll durch Nachahmung des menschlichen Leseverhaltens ein Schritt in Richtung papierloses Büro ausgeführt werden. Exemplarisch werden in ALV Geschäftsbriefe als Domäne untersucht. Teilgebiete innerhalb des ALV-Projekts sind Layoutextraktion, Logical Labeling, Texterkennung und Textanalyse.

Diese Diplomarbeit fällt in den Bereich der Textanalyse. Die Aufgabenstellung bestand darin, mittels der vorkommenden Wörter (im Brieftext) die Art des Briefes sowie erste Hinweise über die Intention des Briefautors zu ermitteln. Derartige Informationen können von anderen Experten zur weiteren Verarbeitung, Verteilung und Archivierung der Briefe genutzt werden. Das innerhalb der Diplomarbeit entwickelte und implementierte INFOCLAS-System versucht deshalb auf der Basis statistischer Verfahren und Methodiken aus dem Information Retrieval folgende Funktionalität bereitzustellen:

i) Extrahierung und Gewichtung von bedeutungstragenden Wörtern;

ii) Ermittelung der Kernaussage (Fokus) eines Geschäftsbriefs;

iii) Klassifizierung eines Geschäftsbriefs in vordefinierte Nachrichtentypen.

Die dafür entwickelten Module Indexierer, Fokussierer und Klassifizierer benutzen — neben Konzepten aus dem Information Retrieval — eine Datenbasis, die eine Sammlung von Geschäftsbriefen enthält, sowie spezifische Wortlisten, die die modellierten Briefklassen repräsentieren. Als weiteres Hilfsmittel dient ein morphologisches Werkzeug zur grammatikalischen Analyse der Wörter. Mit diesen Wissensquellen werden Hypothesen über die Briefklasse und die Kernaussage des Briefinhalts aufgestellt.

Inhaltsverzeichnis

1. Einleitung und Motivation

Mit den Verfahren und Methoden der Künstlichen Intelligenz (KI) wird versucht, Fähigkeiten des Menschen auf dem Computer nachzuahmen, die der Rechner nur schwer und mit viel Aufwand ausführen kann. Neben dem Sehen, Hören und anderen menschlichen Eigenschaften ist das Verstehen von natürlicher Sprache eine dieser Fähigkeiten, die der Mensch mit Abstand besser beherrscht als die Maschine. Probleme, vor denen der Computer steht, sind nicht nur die Vielfalt und Kombinationsmöglichkeiten der Sprache, sondern auch das Verstehen und Interpretieren von Mehrdeutigkeiten, Metaphern, schnellen Kontextwechseln etc. Da für die Erkennung der Bedeutung dieser Sprachkomponenten ein sehr großes Hintergrundwissen (common sense) benötigt wird, welches jedem Menschen durch seine Lebenserfahrung zur Verfügung steht und sehr schnell zugreifbar ist, entstehen für den Rechner Schwierigkeiten bei der Bearbeitung von natürlichsprachlichen Texten. Diese Fülle von Wissen kann einem Rechner nur schwer zugänglich gemacht werden bzw. wenn er sie in seinem Speicher tatsächlich vollständig repräsentieren könnte, wären die Zugriffszeiten durch lange Suchvorgänge sehr groß.

Deshalb wird mit Hilfe von KI-Methoden versucht, Defizite des Computers auszugleichen und ihm damit die Möglichkeiten zu geben, Wörter und Sätze natürlicher Sprache zu interpretieren und zu verarbeiten.

Bei der Analyse von geschriebener natürlicher Sprache entsteht als erstes das Problem der Texterkennung, falls der Text zuvor noch nicht in eine elektronische Form aufbereitet wurde, die der Computer verstehen kann. Hierzu sind intelligente Schnittstellen notwendig. Als Voraussetzung müssen papiergebundene Texte erst mit Hilfe eines Scanners dem Rechner zugänglich gemacht werden.

In einem ersten Verarbeitungsschritt ist es notwendig, daß die einzelnen Zeichen erkannt und zu Wörtern und Sätzen zusammengesetzt werden. Danach kann mit der Interpretation begonnen werden, um die Bedeutung von Wörtern zu ermitteln.

Die Entwicklung von Methoden zur Lösung solcher Aufgabenstellungen, die auf die Domäne der Geschäftsbriefe beschränkt sind, ist das Ziel des ALV-Projekts (**A**utomatisches **L**esen und **V**erstehen) [Dengel et al92a]. Innerhalb dieses Projektes wird versucht, durch Nachahmung des menschlichen Leseverhaltens, Methodiken zu entwickeln, die zur Unterstützung eines papierlosen Büros der Zukunft dienen sollen [Dengel et al92b].

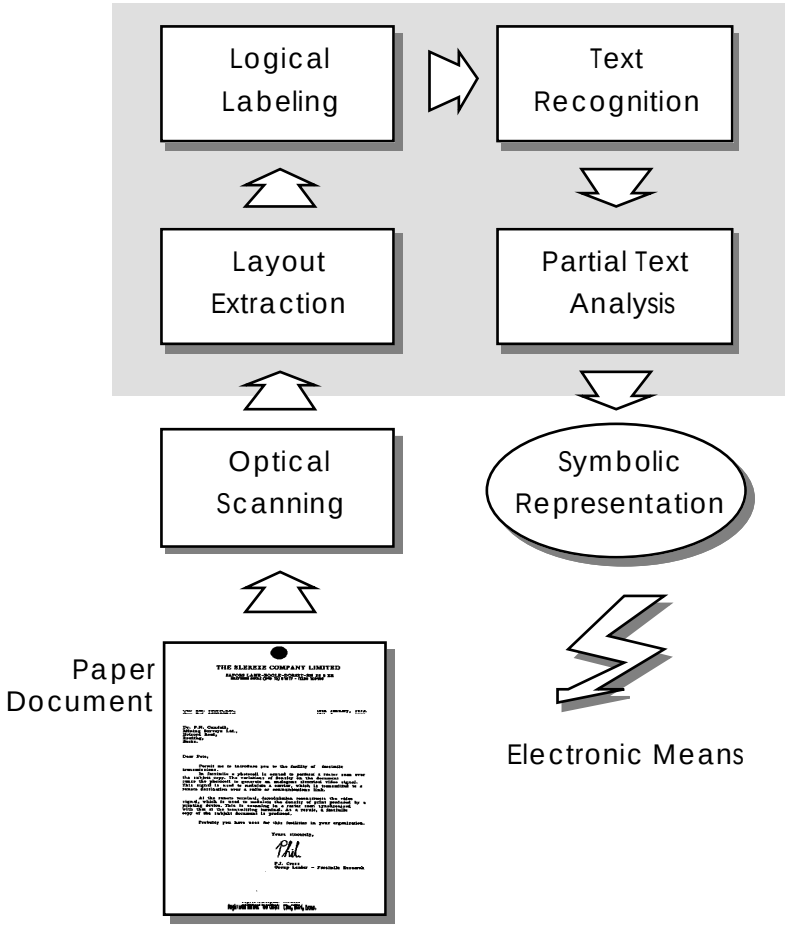


Bild **1.1**: Bearbeitungsmodell des ALV-Projekts

Das Bearbeitungsmodell, welches dem ALV-Projekt zugrunde liegt, wird im Bild 1.1 aufgezeigt. Hier werden die Abfolge der einzelnen Bearbeitungsschritte dargestellt: Layout Extraction, Logical Labeling, Text Recognition und Partial Text Analysis. Hierbei werden mit einem Scanner eingelesene Geschäftsbriefe durch Layout- und Texterkennungsprogramme analysiert und später in ein spezielles Dokumentenmodell transformiert. Ein Ergebnis dieser Transformation ist die Aufteilung des Briefinhaltes in *logische Objekte*. Dieses sind Bereiche innerhalb des Briefes, die in einem bedeutungsmäßigen Zusammenhang (human perceptible meaning) stehen. Beispiele dafür sind: Empfängeradresse, Senderadresse, Betreffteil, Briefrumpf, usw. Exemplarisch zeigt Bild 1.2 die logische und layoutspezifische Strukturierung eines Geschäftsbriefes innerhalb des ALV-Projekts [Bleisinger&Hoch&Dengel91], [Dengel&Hoch92].

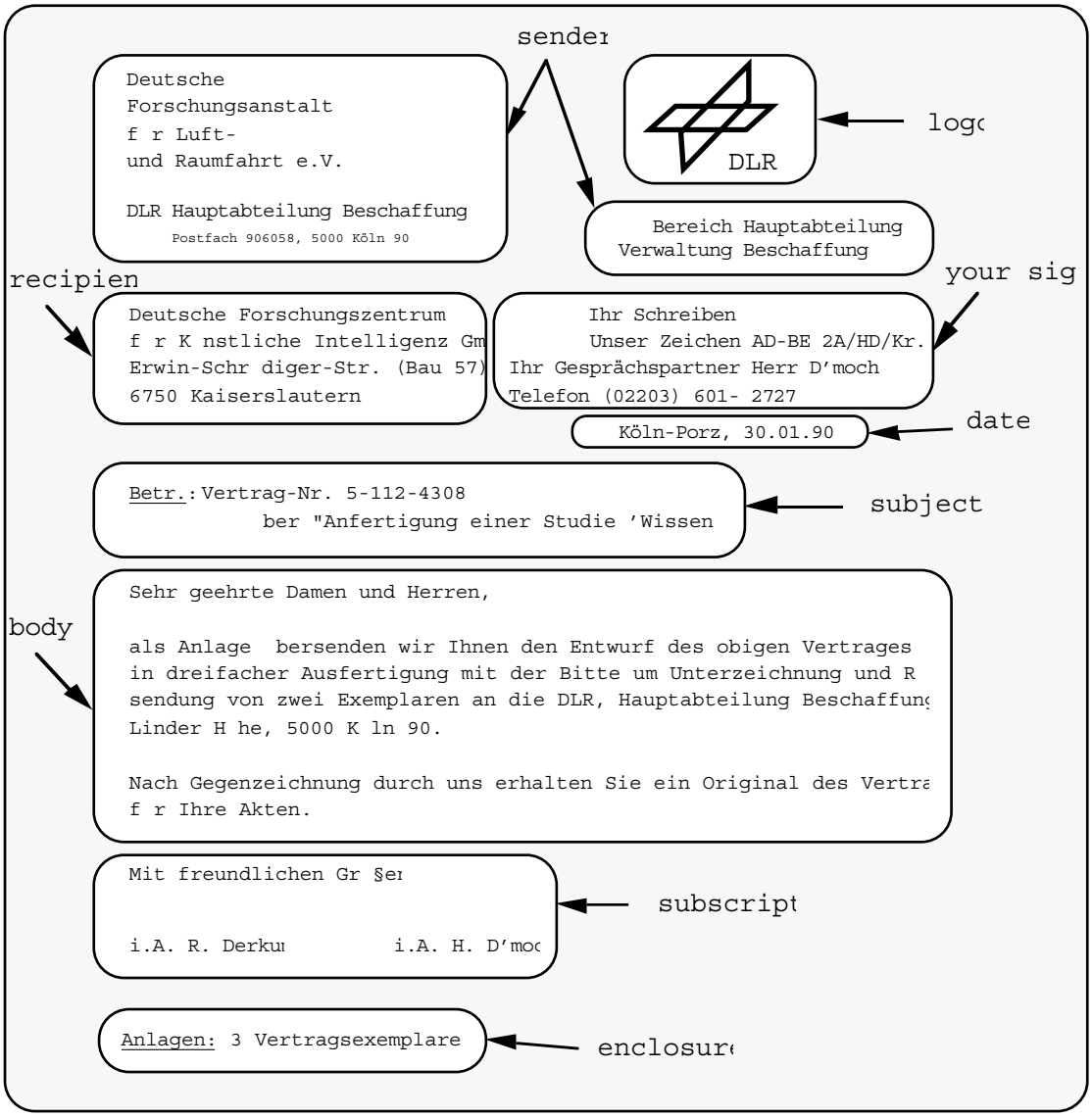


Bild **1.2**: Beispielbrief mit angezeigten logischen Objekten

In einem zweiten Schritt soll nun eine Analyse der Inhalte dieser logischen Objekte erfolgen. Bei der Untersuchung von logischen Objekten wie Empfängeradresse und Senderadresse ist eine natürlichsprachliche Analyse nicht unbedingt notwendig, da diese Teile stark strukturiert sind und keine vollständigen Sätze enthalten. Schwieriger wird es bei dem Versuch, den Sinn und die Bedeutung von logischen Objekten mit ausformulierten Sätzen festzulegen. Die Analyse derartiger Teile des Briefes ist aber besonders wichtig, da hier die Aussage des Briefes verborgen ist.

Solche Aufgaben müßten mit Methoden der natürlichsprachlichen Analyse angegangen werden, um hinter das Geheimnis des Inhaltes zu kommen. Dies ist aber

ohne vorherige Untersuchung der Texte zur Einschränkung des Wissensgebietes, welches zur Analyse gebraucht wird, nur mit großem Aufwand möglich.

Der im ALV-Projekt verfolgte Ansatz orientiert sich an einer Blackboard-Architektur. Dabei greifen mehrere Experten gleichzeitig auf eine Blackboard-Struktur zu, um eigene Ergebnisse darauf abzulegen und temporäre Resultate von anderen Experten als Eingabe aufzunehmen. Hierfür ist nötig, daß jeder Experte ständig auf der Blackboard nachschaut, ob für ihn relevante Zwischenergebnisse vorliegen, welche zur Aktivierung des Experten führen können. Es ist also auch möglich und sinnvoll, daß aufgrund von Veränderungen auf der Blackboard (durch neu erstellte Hypothesen) Experten wiederholt aktiviert werden, um die Analyse des Brieftextes zu verbessern.

Im Rahmen der vorliegenden Diplomarbeit wurde das System *INFOCLAS* entwickelt. INFOCLAS beinhaltet drei Experten zur Textanalyse. Zwei dieser Experten sind der *Fokussierer* und der *Klassifizierer*. Die Aufgabe des Fokussierers ist die Aufstellung von Hypothesen über den möglichen Fokus des Geschäftsbriefes, d.h. den Satz im Brief, der die Kernaussage über den Inhalt des Brieftextes enthält. Der Klassifizierer versucht den aktuell zu bearbeitenden Geschäftsbrief in eine Klasse einzuordnen. Die Klassen, auch *Nachrichtentypen* (*message types*) genannt, sind vordefiniert. Beispiele von Nachrichtentypen sind Angebot, Bestellung, Anfrage usw.

Grundlage für beide Programme ist der, ebenfalls innerhalb der Diplomarbeit entwickelte und als eigenständiger Experte einsetzbare, *Indexierer*. Aufgabengebiet des Indexierers ist es, den einzelnen Wörtern aus einem Brief Gewichte zuzuordnen. Diese Gewichte sollen die Bedeutung dieser Wörter für den Geschäftsbrief widerspiegeln. Mit Hilfe der Wortgewichte und anderem Wissen über die Briefe stellen der Fokussierer und Klassifizierer Hypothesen über den aktuellen Geschäftsbrief auf, die dann von anderen Experten (z. B. einem Inselparsing-Experten) als Basis für ihre Analyse benutzt werden.

2. Grundlagen

2.1. Klassische Information Retrievalsysteme und Methoden zur automatischen Indexierung

Ein Teil der Verfahren, die dieser Diplomarbeit zugrunde liegen, wurden bereits erfolgreich auf dem Gebiet des *Information Retrieval* eingesetzt ([Salton75], [Mresse84], [vanRijsbergen79]). Die Methoden werden hier verwendet, um aus Dokumenttexten die Wörter zu extrahieren, die charakteristisch für diese Dokumente sind. Mit den ermittelten Wörtern, auch *Schlüsselwörter* oder *Deskriptoren* genannt, wurden im weiteren Verlauf der Textverarbeitung die Dokumente repräsentiert. Dabei sollten die Deskriptoren dergestalt sein, daß sie das Dokument von anderen Dokumenten differenzieren. Die Methoden, die dafür eingesetzt werden und ihr Einsatzgebiet, traditionelle Information Retrievalsysteme, sind das Thema der folgenden Abschnitte. In [Salton83] [Salton87] und [Salton89] wird ein guter Überblick über Information Retrievalsysteme und die darin verwendeten Indexierungsverfahren gegeben.

2.1.1. Information Retrievalsysteme

Das Aufgabengebiet von *Information Retrievalsystemen* ist die Repräsentation, Speicherung und Organisation von Informationen und der Zugriff auf diese Daten. Es gibt zwar im Grunde keine Einschränkung, welcher Art diese Informationen sind, aber in den meisten Fällen liegen diese Daten in Form von Texten vor. Beispiele für diese Texte sind Briefe, Dokumente, Bücher, Berichte usw.

Damit vom Benutzer eines Information Retrievalsystems die Informationen auch abgerufen werden können, müssen Zugriffsfunktionen bereit gestellt werden. Diese Zugriffsfunktionen sind nicht von trivialer Natur, da dem anfragenden Benutzer nicht in jedem Fall der genaue Titel, Autor oder das Erscheinungsdatum des Textes bekannt ist. Ebenso ist es möglich, daß er mehrere Dokumente zu einem bestimmten Sachgebiet nachgewiesen haben möchte, oder ihm sind nur ein paar Stichwörter, die sein Interessensgebiet umschreiben, bekannt.

Aus diesem Grund nützt es nicht viel, wenn einfach nur die Titel und Autoren der Texte abgespeichert werden und sie für die Suchstrategie zur Verfügung stehen. Vielmehr müssen auch Informationen aus den Kurzfassungen (abstracts) und/oder direkt aus dem Inhalt verwendet werden, um auch bei unvollständigen Angaben des Anfragenden ein zufriedenstellendes Ergebnis zu liefern.

Die Grundstrukturen des Information Retrieval sollen nun kurz vorgestellt werden.

Jedes Retrievalsystem besteht aus einer Anzahl von Dokumenten (DOKS) und einer Menge von Suchanfragen (FRAGEN). Dazu gehört ein Mechanismus (ÄHNLICH), der aus der Menge der Dokumente diejenigen auswählt, die für die gestellte Suchanfrage relevant erscheinen. Diese Grundstruktur kann folgendermaßen verdeutlicht werden:

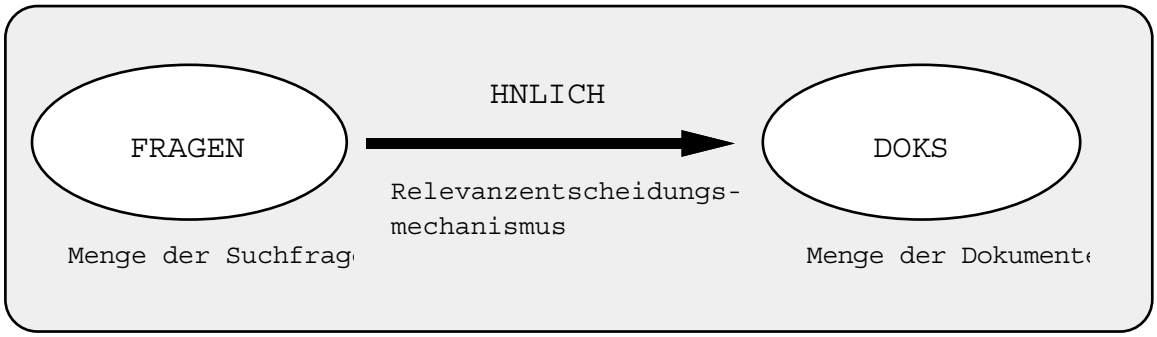


Bild **2.1**: Grundstruktur eines Information Retrievalsystems

Da ein einfacher Abgleich von Suchanfragen mit den Texten wenig sinnvoll erscheint, wird die Grundstruktur um eine Komponente *Indexierungssprache* (*Indexsprache*) erweitert. Diese Indexsprache bildet die Schnittstelle zwischen Suchanfragen und Dokumenten und besteht aus Begriffen, die die Dokumente repräsentieren sollen. In diesem erweiterten Modell eines Information Retrievalsystems werden die Suchanfragen in die Indexsprache (SPRACHE) übersetzt. Ebenfalls wird die Menge der Dokumente mit Hilfe der *Indexierung* dem Entscheidungsprozeß zugänglich gemacht. Mit diesem Prozeß soll die Ähnlichkeit der übersetzten Suchanfragen mit den indexierten Dokumenten ermittelt werden.

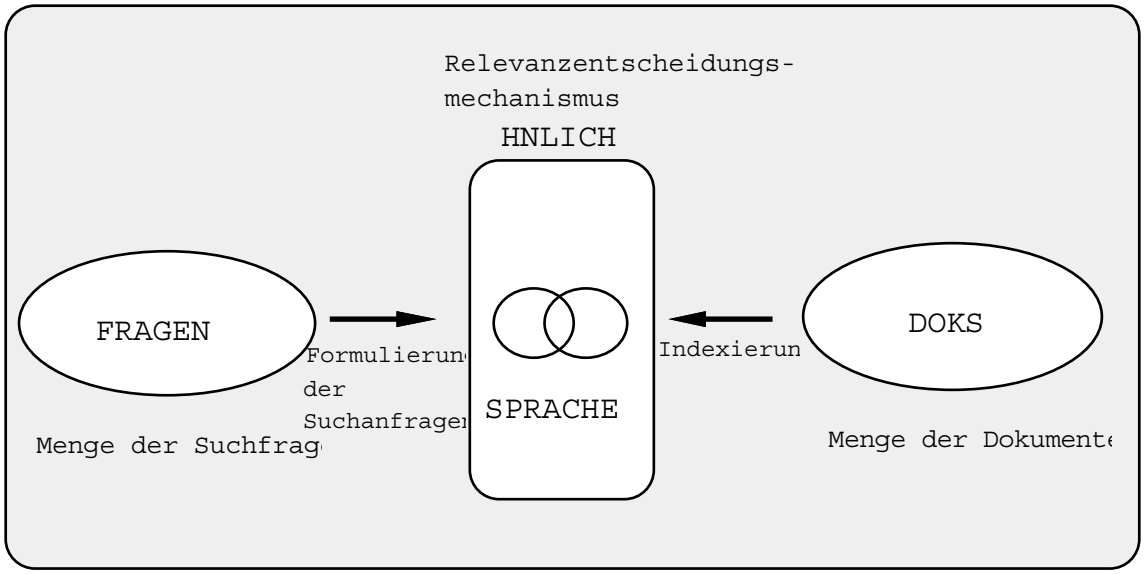


Bild **2.2**: Erweiterte Grundstruktur eines Information Retrievalsystems

Die Indexsprache kann aus Begriffen bestehen, die im voraus festgelegt wurden (*kontrollierte Indexsprache*), oder aus Begriffen, die aus den Suchfragen und Dokumenten direkt ermittelt werden (*unkontrollierte Indexsprache*). Es sind aber auch Mischformen denkbar. Für die Effizienz des Retrievalvorgangs ist die geschickte Auswahl der Begriffe (*Deskriptoren*), die die Dokumente repräsentieren sollen, von entscheidender Bedeutung.

2.1.2. Textanalyse und automatische Indexierung

Die Begriffe, die ein bestimmtes Dokument repräsentieren sollen, kann man in zwei Arten von Informationen unterteilen, in *formale* und *inhaltliche Informationen*. Bei ersteren handelt es sich um Angaben wie Autor, Verlag, Erscheinungsdatum usw., welche sich nicht direkt auf den Dokumenttext beziehen. Dazu im Gegensatz, die inhaltlichen Informationen, die aus dem Dokumenttext direkt entnommen werden und die Thematik des Dokuments beinhalten sollen. Bezieht man sich bei der Indexierung nur auf die formalen Deskriptoren, ist nicht gewährleistet, daß man damit ein ausreichendes Retrievalergebnis erzielt.

Interessanter und wichtiger sind deshalb die inhaltsbezogenen Deskriptoren, mit denen drei Ziele verfolgt werden:

1. Suche nach Dokumenten, die für die Anfrage eines Benutzers relevant sind;
2. Verknüpfung der Dokumente, die thematisch zusammengehören;
3. Bestimmung des Relevanzgrades der gefundenen Dokumente aufgrund der sie repräsentierenden Begriffe.

Um diese Ziele zu erreichen, können verschiedene Arten der Indexierung benutzt werden. Man kann unterscheiden zwischen *manueller, halbautomatischer* und *automatischer Indexierung*. Bei der manuellen Indexierung wird meist eine kontrollierte Indexierungssprache verwendet, im Gegensatz zu meist unkontrollierter Indexsprache bei automatischen Indexierungsverfahren. Ebenso kann bei der Art des Vokabulars unterschieden werden zwischen *singulären* und *kontextbezogenen Deskriptoren*. Singuläre Deskriptoren sind aus einem Begriff bestehende Deskriptoren, kontextbezogene Deskriptoren hingegen setzen sich aus durch spezielle Relationen verbundene Begriffe zusammen, z.B. *Mehrwortbegriffe*. Diese bestehen aus zwei oder mehr Begriffen, die innerhalb eines Dokuments in engerer Beziehung zueinander stehen. Ein Beispiel dafür sind die beiden Begriffe "Information" und "Retrieval" aus denen der Mehrwortbegriff "Information Retrieval" gebildet werden kann (näheres siehe Abschnitt *2.1.4.2. Mehrwortbegriffe*).

Die Bewertung eines Retrievalsystems kann anhand zweier Parameter durchgeführt werden: *Recall* und *Precision*.

Dabei wird unterschieden zwischen verschiedenen Arten von Informationen oder Dokumenten, die im Bezug zum Benutzer und damit zur Suchanfrage stehen:

1. Relevante Informationen oder Dokumente:

   Dies sind Informationen, die der Benutzer mit seiner Suchanfrage vom System ausgegeben bekommen möchte, also Dokumente, die sein Interessensgebiet behandeln.

2. Nachgewiesene oder gefundene Informationen oder Dokumente:

   Dies sind Informationen, die das System aufgrund der Suchanfrage dem Benutzer ausgibt. Diese Informationen oder Dokumente müssen aber nicht immer relevant für den Benutzer sein (Schwächen im Retrievalmechanismus).

Anhand dieser Typen von Informationen können die Parameter Recall und Precision definiert werden.

Mit dem Recall wird der Anteil an relevanten Informationen gemessen, der auf eine Suchanfrage hin nachgewiesen wurde (Anzahl der gefundenen relevanten Dokumente dividiert durch die Anzahl der in der gesamten Dokumentation vorhandenen relevanten Dokumente).

Mit Precision wird verdeutlicht, wieviele der nachgewiesenen Dokumente relevant sind (Anzahl der gefundenen relevanten Dokumente dividiert durch die Anzahl der nachgewiesenen Dokumente).

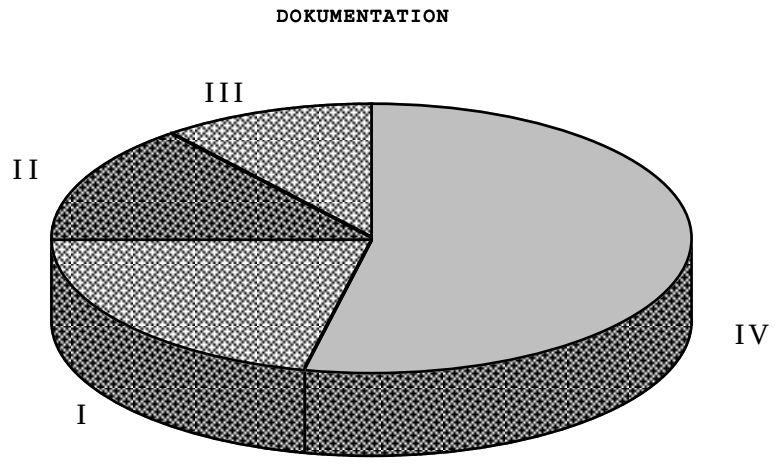


Bild **2.3**: Aufteilung der Dokumentation nach Suchanfragenbearbeitung

In der graphischen Darstellung in Bild 2.3 werden die verschiedenen Teilmengen der Dokumentation nach Bearbeitung einer Suchanfrage angezeigt:

I) Anzahl der nachgewiesenen, nicht relevanten Dokumente;

II) Anzahl der nachgewiesenen, relevanten Dokumente;

III) Anzahl der nicht nachgewiesenen, relevanten Dokumente;

IV) Anzahl der nicht nachgewiesenen, nicht relevanten Dokumente.

Die Flächen I und II enthalten die Dokumente, die auf die Suchanfrage hin ausgegeben werden. Die Dokumente, die auf eine Suchanfrage hin nachgewiesen werden sollten, befinden sich in den Flächen II und III. In einem System, welches einen hohen Recallwert haben soll, muß die Fläche III möglichst klein gehalten werden. In einem System mit hohem Precisionwert sollte die Fläche I auf ein Minimum reduziert werden. Dies bedeutet, daß mit Recall die Eigenschaft des Systems gemessen wird, relevante Dokumente nachzuweisen und mit Precision die Fähigkeit nicht relevante Dokumente zurückzuweisen.

Ein hoher Recallwert wird mit einer *breiten* Indexsprache (*shallow indexing*) erreicht, d.h. es wird mit wenigen allgemeinen Begriffen indexiert. Dies hat aber eine Reduzierung der Precision zur Folge. Ein hoher Precisionwert wird mit *tiefer* Indexierung (*deep indexing*) erreicht, d.h. mit spezifischen und präzisen Deskriptoren, welches sich aber negativ auf den Recallwert auswirkt. Das Ziel eine Indexierungssprache mit gleichzeitig hohem Recallwert und Precisionwert festzulegen, ist nur schwer erreichbar und erfordert eine exakte Auswahl der Deskriptoren.

2.1.3. Automatische Generierung von Deskriptoren und Deskriptorgewichten

2.1.3.1. Relative Häufigkeit von Begriffen

Indexieren bedeutet, jedem gespeicherten Dokument Deskriptoren zuzuordnen, die deren Inhalt repräsentieren sollen. Weiterhin sollten die Deskriptoren einen Wert ihre Bedeutung für das Dokument zugeordnet bekommen (*Gewichtung*). Grundlegende Überlegung bei der automatischen Indexierung ist, daß Wörter, die häufig auftreten eine höhere Bedeutsamkeit besitzen als Wörter, welche weniger oft im Text erscheinen. Dies ist aber, wie im weiteren zu sehen ist, eine zu einfache Sicht, die noch durch erweiternde Konzepte verbessert werden muß.

Betrachtet man nämlich die Wörter mit sehr hoher Frequenz, erkennt man, daß diese sogenannten *Funktionswörter* nichts mit dem Inhalt und der Thematik des Textes zu tun haben. Diese Wörter stammen aus Wortgruppen wie Konjunktionen, Artikel, Partikel, Pronomen usw. Beispiele sind: “und“, “oder“, “der“, “die“, “das“ usw. Es wäre also sinnlos, diese Wörter als Deskriptoren auszuwählen, gerade unter dem Aspekt, daß sie in jedem Text erscheinen und damit nicht charakteristisch für ein bestimmten

Textinhalt sein können. Ebenso sind Wörter mit einer sehr geringen Häufigkeit keine guten Deskriptoren, da sie zu selten in der Dokumentation auftreten, um das Retrievalergebnis entscheidend zu beeinflussen.

Aus diesen Überlegungen ergibt sich folgendes einfache Verfahren:

1. Errechne für jedes Dokument die Häufigkeit eines jeden Begriffs innerhalb des Dokuments. Sie wird als Häufigkeit des Begriffs k im Dokument i mit $FREQ_{ik}$ bezeichnet.
2. Bestimme für jeden Begriff die Häufigkeit in der ganzen Dokumentation, $TOTFREQ_k$ bezeichnet, durch Aufsummieren der Häufigkeiten eines einzelnen Begriffs k über alle Dokumente.

   $$TOTFREQ_k = \sum_{i=1}^{n} FREQ_{ik}$$

3. Ordne die Begriffe nach abnehmender Häufigkeit. Eliminiere alle Wörter oberhalb eines angemessenen oberen Schwellwertes. Dadurch erfolgt die Abspaltung der hochfrequenten Funktionswörter.
4. Eliminiere alle Wörter unterhalb eines unteren Schwellwertes und schließe damit die zu wenig auftretenden, niederfrequenten Begriffe aus.
5. Alle übrigen Wörter werden als Deskriptoren eingesetzt.

In diesem sehr einfachen Verfahren gibt es aber noch eine Anzahl von Nachteilen:

i) Nicht alle Hochfrequenzbegriffe müssen Funktionswörter sein und sollten deshalb berücksichtigt werden, um damit einen Verlust beim Recall zu vermeiden.

ii) Ebenso dürfen nicht alle Niederfrequenzbegriffe eliminiert werden, da dies zu einem Verlust bei der Precision führen kann.

iii) Die Ermittlung von angemessenen Schwellwerten ist nicht ohne Probleme zu bewältigen.

iv) Die Bewertung der Deskriptoren durch absolute Häufigkeiten ist ein Problem, da ein Deskriptor in Retrievalsystemen zwischen einzelnen Dokumenten unterscheiden soll, dies aber kann durch eine einfache Bewertung nach absoluter Häufigkeit nicht immer zufriedenstellend geleistet werden. Zum Beispiel in einer Dokumentation über Computer wird der Begriff *Computer* oft in allen Dokumenten auftreten und kann somit zur Unterscheidung der Dokumente nicht nützlich sein.

Aus diesen Gründen sind weitere Konzepte und Bewertungsfunktionen bei der automatischen Indexierung notwendig, die in den folgenden Abschnitten beschrieben werden.

2.1.3.2. Inverse Dokumenthäufigkeit

Zur Verbesserung der Bewertung der Begriffe kann das Konzept der inversen Dokumenthäufigkeit eingesetzt werden. Bei diesem Konzept wird angenommen, daß die Bedeutsamkeit eines Begriffs k mit der Häufigkeit dieses Begriffs im Dokument i ($FREQ_{ik}$) steigt und umgekehrt proportional zur Gesamtanzahl der Dokumente ist, denen der Begriff k zugeordnet wurde. Dabei zählt man die Anzahl der Dokumente, die der Begriff k als Deskriptor zugeordnet ist ($DOKFREQ_k$).

Die *Gewichtsfunktion*, die daraus abgeleitet werden kann, wird nach Formel (G1) berechnet, wobei n die Anzahl der Dokumente in der Dokumentation ist.

$$GEWICHT_{ik} = FREQ_{ik} * [\log_2 (n) - \log_2 (DOKFREQ_k) + 1] \quad (G1)$$

Durch diese Gewichtsfunktion wird den Begriffen, die in wenigen Dokumenten vorkommen, eine höhere Bedeutsamkeit beigemessen, als solchen, die in vielen Dokumenten vorkommen. Die Bedeutung steigt also, wenn ein Begriff häufig in wenigen Dokumenten auftritt.

2.1.3.3. Informationswert und Ballast

Eine weitere Möglichkeit zur Generierung von Deskriptorgewichten hat ihren Ursprung in der Informationstheorie. Dort nimmt man an, daß ein Begriff genau dann einen hohen *Informationsgehalt* besitzt, wenn die relative Wahrscheinlichkeit des Auftretens dieses Begriffes gering ist. Dies spiegelt sich in der Formel (1) wider, wobei p die relative Häufigkeit des Begriffes darstellt.

$$INFORMATION = - \log_2 p \quad (1)$$

Im Gegensatz dazu kann man den *Ballast* $NOISE_k$ eines Begriffs k ermitteln. Dieser Wert ist hoch, wenn der Begriff auf alle Dokumente gleich verteilt ist, also überall gleich oft vorkommt. Er ist 0, wenn der Begriff nur in einem Dokument vorhanden ist.

Dies wird durch mit der Gleichung (2) ermittelt, wobei n für die Anzahl der Dokumente in der Dokumentation steht.

$$NOISE_k = \sum_{i=1}^{n} \frac{FREQ_{ik}}{TOTFREQ_k} * \log_2 \frac{TOTFREQ_k}{FREQ_{ik}} \qquad (2)$$

Nach diesen Überlegungen läßt sich nun der *Informationswert* eines Begriffes k festlegen durch die Gleichung (3).

$$SIGNAL_k = \log_2 (TOTFREQ_k) - NOISE_k \qquad (3)$$

Die Gleichung (3) gibt denjenigen Begriffen einen hohen Informationswert, die einen geringen Ballast und TOTFREQ besitzen. Wenn man die relative Häufigkeit eines Begriffes in einem Dokument i dazu nimmt, kann man eine Gewichtsfunktion (G2) festlegen.

$$GEWICHT_{ik} = FREQ_{ik} * SIGNAL_k \qquad (G2)$$

2.1.3.4. Diskriminanzwert

Die bisher vorgestellten Gewichtungsfunktionen für Deskriptoren basierten auf der *relativen Häufigkeit* eines Begriffes. Eine andere Möglichkeit besteht in der Ermittlung des *Diskriminanzwertes*. Hierbei wird die Fähigkeit eines Begriffes ermittelt, zwischen Dokumenten zu unterscheiden.

Dafür wird eine Funktion (ÄHNLICH (D_i,D_j)) benötigt, die die Ähnlichkeit der Dokumente D_i und D_j ermittelt. Dabei werden die Deskriptoren, die den Dokumenten D_i,D_j zugeordnet sind, verglichen. Stimmen sie gänzlich überein, erhält man den Ähnlichkeitsfaktor 1. Besteht überhaupt keine Übereinstimmung, bekommt man den Wert 0 geliefert. Partielle Übereinstimmungen liegen zwischen 0 und 1. Wie konkrete Ähnlichkeitsfunktionen aussehen können wird später beschrieben.

Sind nun alle Ähnlichkeitsfaktoren für alle Kombinationen von Dokumentenpaaren bestimmt, so ist es möglich, die Durchschnittsähnlichkeit mit der Gleichung (4) zu berechnen.

$$\text{DURCHSCHNITTSÄHNLICHKEIT} = \text{KONSTANTE} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ (i \neq j)}}^{n} \text{ÄHNLICH}(D_i, D_j) \qquad (4)$$

Beispiel für die Konstante: KONSTANTE = $1/n * (n-1)$

Der ermittelte Wert aus Gleichung (4) wird auch die *Dichte des Dokumentenraums* genannt.

Eine effizientere Möglichkeit zur Bestimmung der Dichte besteht in der Einführung eines künstlichen durchschnittlichen Dokuments D´ (*Zentroid*). Dieser Zentroid enthält alle Begriffe mit durchschnittlicher Häufigkeit. Dabei ist die durchschnittliche Häufigkeit des Begriffs k definiert durch die Formel (5):

$$(\text{DURCHSCHNITTSFREQ})_k = \frac{1}{N} * \sum_{i=1}^{n} \text{FREQ}_{ik} \qquad (5)$$

Die Dichte des Dokumentenraums ergibt sich dann aus der Summe der Ähnlichkeiten eines Dokuments i mit dem Zentroid D´ (Gleichung 6):

$$\text{DURÄHN} = \text{KONSTANTE} \sum_{i=1}^{n} \text{ÄHNLICH}\,(D', D_i) \qquad (6)$$

Wird nun die Dichte des Dokumentenraums ohne den Begriff k ermittelt (DURÄHN_k), d.h. wird er nicht mehr als Deskriptor benutzt, können folgende Fälle auftreten:

i) Die Ähnlichkeit zwischen den Dokumenten wird reduziert; d.h. dieser Begriff ist kein guter Deskriptor, da er die Dichte des Dokumentenraums vergrößert, und nicht zur besseren Unterscheidung der Dokumente dient.

ii) Die Ähnlichkeit nimmt zu. Daraus folgt, der Begriff k hat für einige Dokumente ein hohes Gewicht und für die restlichen eine geringere Bedeutung. Diese macht ihn zu einem guten Deskriptor, da er zur Differenzierung der Dokumente beiträgt.

Der Diskriminanzwert DISKWERT_k für einen Begriff k ist somit festlegbar durch Gleichung (7):

$$\text{DISKWERT}_k = (\text{DURÄHN}_k) - \text{DURÄHN} \qquad (7)$$

Nach dieser Gleichung kann man drei Kategorien von Diskriminatoren unterscheiden:

1) Gute Diskriminatoren mit positivem $DISKWERT_k$.
   (Reduzieren der Dichte des Dokumentenraums)
2) Indifferente Diskriminatoren mit Werten nahe 0.
   (Keine Veränderung der Dichte und damit kaum einsatzfähig als Deskriptor)
3) Schwache Diskriminatoren mit negativem $DISKWERT_k$.
   (Die Dokumente werden ähnlicher. Diese Begriffe sind also nicht geeignet als Deskriptoren)

Der Diskriminationswert kann auch zur Berechnung einer Gewichtsfunktion (G3) benutzt werden, mit dem für jeden Begriff innerhalb eines Dokuments sein Bedeutungswert errechnet wird.

$$GEWICHT_{ik} = FREQ_{ik} * DISKWERT_k \qquad (G3)$$

2.1.3.5. Ähnlichkeitsfunktionen

Abschließend sollen drei mathematische Funktionen zur Bestimmung der Ähnlichkeit zwischen Deskriptoren vorgestellt werden [Salton 83].

Wie schon erwähnt, werden Ähnlichkeitsfunktionen dazu benutzt, den Grad der Übereinstimmungen zwischen Deskriptorvektoren und damit den von ihnen repräsentierten Dokumenten zu ermitteln. Dabei können die Vektoren aus Deskriptorgewichten bestehen, die nur 1 oder 0 annehmen können (*binäres Indexsystem*), oder solchen, die auch Werte zwischen 0 und 1 bekommen können (*gewichtetes Indexsystem*). In beiden Fällen können die angegebenen Ähnlichkeitsfunktionen angewendet werden. Allen ist gemeinsam, daß der Ähnlichkeitswert ansteigt, wenn die Anzahl der gemeinsamen Merkmale zunimmt.

i) Dicekoeffizient

$$ÄHN_1(DOK_i, DOK_j) = \frac{2 * [\sum_{k=1}^{t} (TERM_{ik} * TERM_{jk})]}{\sum_{k=1}^{t} TERM_{ik} + \sum_{k=1}^{t} TERM_{jk}}$$

ii) Jaccardkoeffizient

$$ÄHN_2(DOK_i,DOK_j) = \frac{\sum_{k=1}^{t}(TERM_{ik} * TERM_{jk})}{\sum_{k=1}^{t}TERM_{ik} + \sum_{k=1}^{t}TERM_{jk} - \sum_{k=1}^{t}(TERM_{ik} * TERM_{jk})}$$

iii) Cosinuskoeffizient

$$ÄHN_3(DOK_i,DOK_j) = \frac{\sum_{k=1}^{t}(TERM_{ik} * TERM_{jk})}{\sqrt{\sum_{k=1}^{t}(TERM_{ik})^2 * \sum_{k=1}^{t}(TERM_{ik})^2}}$$

(In allen Funktionen bedeutet $TERM_{ik}$ das Gewicht des Deskriptors k für das Dokument i)

Beim Cosinuskoeffizienten ist zu beachten, daß die Kennziffer für den Winkel zwischen den zwei Deskriptorvektoren ermittelt wird, da diese als Vektoren eines t-dimensionalen Vektorraums angesehen werden können.

2.1.4. Begriffsassoziationsverfahren (assoziative Indexierung)

Beim ersten einfachen Verfahren aus Abschnitt 2.1.3.1. konnten einige Mängel erkannt werden, die nun durch erweiternde Methoden verbessert werden sollen.

Etwa 40-50% der Begriffe in Texten sind hochfrequente Funktionswörter, die schlechte Diskriminatoren und damit als Deskriptoren ungeeignet sind. Diese können mit Hilfe einer *Stoppwortliste* erkannt werden. Sie kommen als Deskriptoren dann nicht mehr in Frage. In dieser Stoppliste stehen Wortformen wie Pronomen, Artikel, Partikel, Konjunktionen etc., die oft benutzt werden, aber nichts mit dem eigentlichen Sinn des Textes zu tun haben. Diese Stoppwortliste muß natürlich für jede Sprache (Englisch, Deutsch usw.) neu erstellt und in das System eingebunden werden.

Eine weitere Methode zur Verbesserung der Deskriptorbestimmung ist die Reduzierung der Begriffe von den Vollformen auf ihre Wortstämme bzw. Grundformen. Dabei werden Wörter, die aus dem gleichen Wortstamm gebildet wurden, durch Entfernung von Wortsuffixen auf diesen Wortstamm reduziert. Zum Beispiel: Arbeiter, arbeiten, arbeitslos werden auf den gemeinsamenWortstamm "arbeit" verkürzt.

Im Englischen kann dies mit Hilfe einer *Suffixliste* geschehen, welche die gängigsten Endungen enthält. Weiterhin müssen eine Anzahl von Kontextregeln beachtet werden, um eine korrekte Reduzierung zu erhalten. Eine solche Wortreduktion ist in flexionsreichen Sprachen wie Deutsch oder Französisch schwieriger. Hier müssen linguistische Regeln beachtet werden, die den Vorgang der Reduzierung auf die Grundformen erschweren. Die Erweiterung des Systems um eine Komponente mit *morphologischer Analyse* erscheint im Deutschen damit als sinnvoll.

Nun werden die Begriffe nicht mehr einzeln betrachtet, sondern nur noch ihre Wortstämme, was zu einer deutlichen Verkleinerung der Anzahl der Begriffe führt. Diese Wortstämme werden als Deskriptoren in die *Deskriptorenvektoren* der Dokumente eingetragen. Zuvor sollte mit der inversen Dokumenthäufigkeit und den Diskriminanzwerten die Bedeutung und Gewichtung dieser Deskriptoren bestimmt werden. Diese ergeben dann ein *gewichtetes Indexsystem* mit gewichteten Deskriptoren. Werden keine Gewichte verteilt, sondern nur angegeben, ob ein Deskriptor für ein Dokument gültig ist oder nicht, spricht man von einem *binären Indexsystem*.

Ein Beispiel für ein Deskriptorvektor zeigt die Gleichung (8), wobei:

D_i = Dokument, d_{ij} = Gewicht des j-sten Deskriptors und n = Anzahl der Dokumente in der Dokumentation.

$$D_i = < d_{i1}, d_{i2}, ..., d_{in} > \quad (8)$$

2.1.4.1. Thesaurus

Wie schon früher erwähnt, kann es zu Verlusten bei Recall und Precision kommen, wenn einfach Hoch- und Niederfrequenzbegriffe eliminiert werden. Deshalb sollten Verfahren, die sich auf Begriffsassoziationen beziehen, eingesetzt werden. Mit diesen Verfahren soll das Bedeutungsfeld einzelner Begriffe präzisiert oder verallgemeinert werden.

Eine Methode ist es, beim Indexieren das Hilfsmittel *Thesaurus* einzusetzen. In einem Thesaurus werden Begriffe klassifiziert und in bestimmte Klassen oder Kategorien eingetragen. In diesen Klassen stehen ähnliche oder zueinander in Beziehung stehende Wörter (*Synonyme*). Bei der Thesauruskonstruktion sollten folgende Regeln beachtet werden:

i) Es sollten nur Begriffe enthalten sein, die für das Sachgebiet von Relevanz sind;

ii) Bei mehrdeutigen Begriffen sollten nur die Bedeutungen gespeichert werden, die für das Sachgebiet von Interesse sind;

iii) Es soll eine gewisse Gleichmäßigkeit der Häufigkeit des Auftretens von Begriffen in den Dokumenten gewährleistet sein, die in eine gemeinsame Klasse

gespeichert werden. Damit sollte bei speziellen Begriffen ein nicht zu allgemeines Ergebnis zurück gemeldet werden.

Thesauri können manuell oder automatisch erzeugt werden. Dabei wird auf die Dokumentvektoren zurückgegriffen. Aus der Matrix der Dokumentvektoren für die gesamte Dokumentation (Bild 2.4) ist auch die Gewichtung des einzelnen Begriffs T_k für alle Dokumente ablesbar.

D	T_1	T_2	...	T_t
D_1	d_{11}	d_{12}	...	d_{1t}
D_2	d_{21}	d_{22}	...	d_{2t}
:	:	:		:
:	:	:		:
D_n	d_{n1}	d_{n2}	...	d_{nt}

Bild **2.4**: Matrix der Dokumentenvektoren

Dabei sind die Spalten der Dokumentvektormatrix:

$$Term_k = (t_{1k}, t_{2k}, ..., t_{nk}) \quad (9)$$

Werden diese paarweise mit einer Ähnlichkeitsfunktion verglichen, wie vorher die Dokumentvektoren bei der Ermittlung der Dichte des Dokumentenraums, so ergibt sich eine Begriffsassoziationsmatrix T mit der Ähnlichkeitsfunktion s.

T	T_1	T_2	...	T_t
T_1	$s(T_1,T_1)$	$s(T_1,T_2)$	...	$s(T_1,T_t)$
T_2	$s(T_2,T_1)$	$s(T_2,T_2)$	...	$s(T_2,T_t)$
:	:	:		:
:	:	:		:
T_t	$s(T_t,T_1)$	$s(T_t,T_2)$	...	$s(T_t,T_t)$

Bild **2.5**: Begriffsassoziationsmatrix

z.B. durch die Funktion (Formel 10) für *s* aus Bild 2.5:

$$\text{ÄHNLICH}(\text{TERM}_k, \text{TERM}_h) = \sum_{i=1}^{n} t_{ik}\, t_{ih} \qquad (10)$$

Durch automatische Klassifikationsmethoden (*Clusterbildung*) können nun die Begriffe in Gruppen zusammengefaßt werden, die dann die Thesaurusklassen bilden. In diesen Klassen befinden sich diejenigen Begriffe vereint, die einander ähnlich sind und als Synonyme gelten können.

Um diese Klassen zu bilden, gibt es zwei Vorgehensweisen. Entweder man geht von den einzelnen Deskriptoren aus und bildet aus ihnen nach und nach die einzelnen Klassen, oder man legt vorher Begriffsklassen fest und verfeinert diese im Laufe des Verfahrens immer mehr bis zu den fertigen Thesaurusklassen.

Ein Beispielverfahren für den ersten Fall ist das *single-linkage* Verfahren.

1. Zu jedem Begriffspaar (TERM_i, TERM_j), dessen Ähnlichkeit einen bestimmten Schwellenwert überschreitet, wird ein dritter Begriff TERM_k hinzugefügt; zwischen jedem Element des Ausgangsbegriffpaares und dem TERM_k wird ein Ähnlichkeitswert berechnet; der neue Begriff wird dieser Begriffsklasse hinzugefügt, wenn der Ähnlichkeitswert mindestens einen der Ähnlichkeitswerte der Ausgangsbegriffe überschreitet.

2. Dieses Verfahren wird dann für Tripel, Quadrupel usw. fortgesetzt, wobei jedesmal ein neuer Begriff der Ausgangsklasse hinzugefügt wird, wenn die Ähnlichkeit dieses Begriffs mit einem Ausgangsbegriff des Schwellenwerts überschreitet.

Thesauri werden dafür verwendet, das Indexierungsvokabular allgemeiner zu gestalten. Die Idee dabei ist es, nicht mit einzelnen Begriffen, sondern mit den Thesaurusklassen als Deskriptoren zu indexieren. Dieses erscheint besonders günstig bei Niederfrequenzbegriffen, die dadurch, obwohl sie keinen eigenen guten Diskriminanzwert besitzen, doch dem Retrievalprozeß zur Verfügung stehen können.

Da die Begriffe sehr selten in den einzelnen Dokumenten auftreten, ist es schwer, eine automatische Thesauruskonstruktion durchzuführen. Es kann deshalb von Vorteil sein, einen manuell erstellten Thesaurus zu verwenden.

2.1.4.2. Mehrwortbegriffe

Ein weiteres Verfahren, das mit Begriffsassoziationen arbeitet, ist die Generierung von *Mehrwortbegriffen*. Mit diesem Verfahren soll die Spezialität einzelner Begriffe erhöht werden. Dies ist natürlich nur sinnvoll bei Begriffen mit einer hohen Frequenz, da bei anderen Wörtern, mit niedriger Häufigkeitsfrequenz, eine Spezialität genügend vorhanden ist und durch Mehrwortbegriffsbildung zu stark verstärkt würde.

Die Mehrwortbegriffe können mit Hilfe von syntaktischen und/oder semantischen Analyseverfahren aus dem Bereich der Erkennung von natürlicher Sprache generiert werden. Dies sind aber meist aufwendige Verfahren. Deshalb wird hier ein weitaus einfacheres Verfahren vorgestellt, welches folgende Kriterien berücksichtigt:

1. Die einzelnen Komponenten des Mehrwortbegriffs sollten in einem Sinnzusammenhang in dem Dokument stehen, dem der Mehrwortbegriff als Deskriptor zugeordnet werden soll.
2. Die Mehrwortbegriffe sollten inhaltlich eher eine allgemeine Bedeutung besitzen; d.h. ihre Dokumenthäufigkeit sollte ausreichend hoch sein.

In dem Verfahren werden Mehrwortbegriffe aus zwei verschiedenen Wortstämmen gebildet, wovon einer davon eine hohe Dokumenthäufigkeit besitzen sollte. Die Wortstämme des zu bildenden Mehrwortbegriffs sollten nicht in der Stoppwortliste auftreten. Weiterhin ist auch sinnvoll, wenn der Abstand von Wortstämmen des Mehrwortbegriffs innerhalb des Dokuments mit berücksichtigt wird und er einen gewissen Schwellenwert nicht überschreitet. Die beiden Komponenten des Mehrwortbegriffes sollten dabei in einem Satz innerhalb eines Dokuments stehen.

Verfahren zur Generierung von Mehrwortbegriffen:

1. Beginne mit dem Text des Dokuments; benutze eine Stoppwortliste, um die allgemeinen Funktionswörter zu entfernen; generiere Wortstämme mit einer Suffixanalyse (bzw. morphologische Analyse);
2. Bilde aus den einzelnen Paaren von Wortstämmen Mehrwortbegriffe, wobei zu beachten ist, daß die Distanz zwischen den einzelnen Komponenten nicht größer als n Wörter ist und daß mindestens eine der Komponenten ein Hochfrequenzbegriff ist; mit diesem Schwellenwert kann die Zahl der generierten Mehrwortbegriffe auf eine handhabbare Größe begrenzt werden;
3. Mehrwortbegriffe mit identischen Komponenten und doppelt erzeugte Mehrwortbegriffe werden eliminiert;
4. Den Mehrwortbegriffen werden Gewichte in Abhängigkeit der Gewichte der einzelnen Mehrwortbegriffskomponenten zugewiesen; bei binärer Gewichtung

mit 0 und 1 ergibt sich das Gewicht des Mehrwortbegriffs aus der Multiplikation der Gewichte der einzelnen Komponenten.

2.1.5. Verfahren zur automatischen Indexierung

Mit obigen Begriffsassoziationsverfahren (Abschnitt 2.1.4.) läßt sich ein verbessertes Verfahren zur automatischen Indexierung angeben [Salton71]:

I. Nehme den Dokumenttext und berechne die einzelnen Worthäufigkeiten;

II. Benutze eine Stoppwortliste, um unerwünschte Wörter zu entfernen und generiere mit einem Suffixalgorithmus Wortstämme (morphologische Analyse im Deutschen);

III. Berechne Begriffsgewichte $TERM_k$ von DOK_i als Funktionen von $FREQ_{ik}$, $DOKFREQ_k$ und $DISKWERT_k$;

IV. Bilde nach dem Prinzip des gemeinsamen Auftretens im Ursprungssatz Mehrwortbegriffe aus Hochfrequenzbegriffen, die keine befriedigenden Begriffsgewichte aufweisen;

V. Bilde mit dem Single-Linkage-Verfahren Thesaurusklassen aus Niederfrequenzbegriffen, die keine befriedigenden Begriffsgewichte aufweisen;

VI. Bilde aus den verbleibenden Einzel- und Mehrwortbegriffen sowie den Thesaurusklassen, zusammen mit den entsprechenden Gewichten, die Dokumentvektoren.

2.1.6. Theoretische und erweiternde Verfahren

2.1.6.1. Fragmentkodierung

Da die Anzahl der Deskriptoren für eine Dokumentation sehr groß sein kann, wird dadurch die Speicherung und Pflege der Deskriptorvektoren sehr umfangreich. Um diesen Aufwand zu reduzieren, kann man sich eines Verfahrens bedienen, welches sich auf die Informationstheorie stützt und auf dem Unterschied zwischen Deskriptorhäufigkeiten basiert.

Die Idee der *Fragmentkodierung* besteht darin, die einzelnen Deskriptoren durch eine geringe Anzahl künstlicher Zeichenketten, den Fragmenten, zu ersetzen, welche in etwa die gleiche Häufigkeitsverteilung besitzen wie die Deskriptoren.

Zur Generierung der Fragmente kann man zwei Methoden verwenden.

i) Die *Agglomerationsmethode*, bei der durch sukzessive Erweiterung aus einzelnen Buchstaben die Fragmente gebildet werden.

oder die später näher vorgestellte:

ii) Die *Verkürzungsmethode*, bei der die vollständigen Wörter oder Strings durch Verkürzung bis auf die einzelnen Buchstaben verkleinert werden und dadurch die Fragmente entstehen.

Verkürzungsmethode:

1.) Die Wörter eines Textes werden in alle möglichen Fragmente zerlegt und in einem Baum strukturiert (vgl. Bild 2.6.);

2.) Die Häufigkeiten werden für alle entstehenden Fragmente ermittelt und den einzelnen Fragmenten zugeordnet;

3.) Sei i die Länge des längsten Wortfragments. Von allen Fragmenten der Länge i wähle das aus, dessen Häufigkeit einen Schwellenwert am geringsten überschreitet. Wenn kein Fragment dieser Voraussetzung entspricht, wird i um 1 verkleinert und dies wird so fortgesetzt bis i = 1 ist;

4.) Wenn ein Fragment in die Menge der Fragmente aufgenommen wurde, werden die Häufigkeiten der Söhne im Baum um die Häufigkeit des ausgewählten Fragments verkleinert;
Beispiel: Wird "at" ausgewählt, werden die Häufigkeiten von "a" und "t" um 100 reduziert.

5.) Die Größe der *Fragmentmenge* sollte auf eine Gesamtzahl der Größenordnung 2^n reduziert werden durch Elimination redundanter Fragmente, die schon in größeren Fragmenten enthalten sind (z.B. 256, um mit einem 8-Bit Kode kodieren zu können). Es sollten aber alle Einzelbuchstaben vorhanden sein.

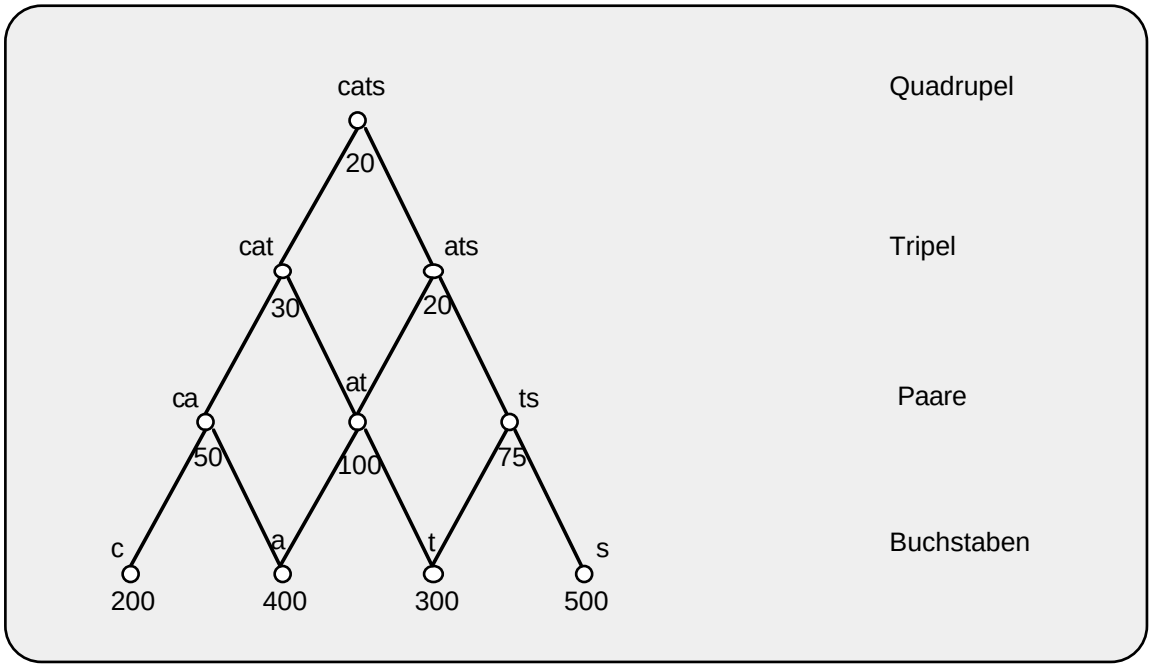


Bild **2.6:** Fragmentbaum für das Wort "Cats"

Diese Methode der Fragmentierung kann auch erweitert werden durch die Berücksichtigung von Leerzeichen, um damit *Mehrwortbegriffsfragmente* zu bilden (z.B. "in der", "zu dem", "auf der").

Nun können die Deskriptoren durch die ermittelten Fragmente ersetzt werden, zum Beispiel mit dem "*longest matching*" Verfahren.Bei diesem Verfahren wird von links nach rechts das größte Fragment im Wort gesucht, was dann diesen Teil des Deskriptors ersetzt. Mit dem Rest des Wortes wird weiter so verfahren, bis kein Rest mehr vorhanden ist. Mit dieser Methode muß nicht immer die effizienteste Kodierung gefunden werden, aber es ist ein schnelles und einfaches Verfahren.

Probleme können durch die Ersetzung der Deskriptoren durch Fragmente beim Einsatz in Retrievalsystemen kommen, da manchmal Doppeldeutigkeiten entstehen und falsche Dokumente nachgewiesen werden.

2.1.6.2. Linguistische Verfahren

Wie früher schon erwähnt, ist ohne eine syntaktische Analyse des Textes die Generierung von Mehrwortbegriffen nicht präzise genug. Um die Erkennung von Mehrwortbegriffen zu verbessern, werden Verfahren entwickelt, die auf linguistischer Basis arbeiten.

Das erste Verfahren verwendet kontextfreie *Phrasenstrukturgrammatiken*. Mit Hilfe dieser Phrasenstrukturgrammatiken wird für jeden Satz eines Dokuments ein Strukturbaum erstellt.

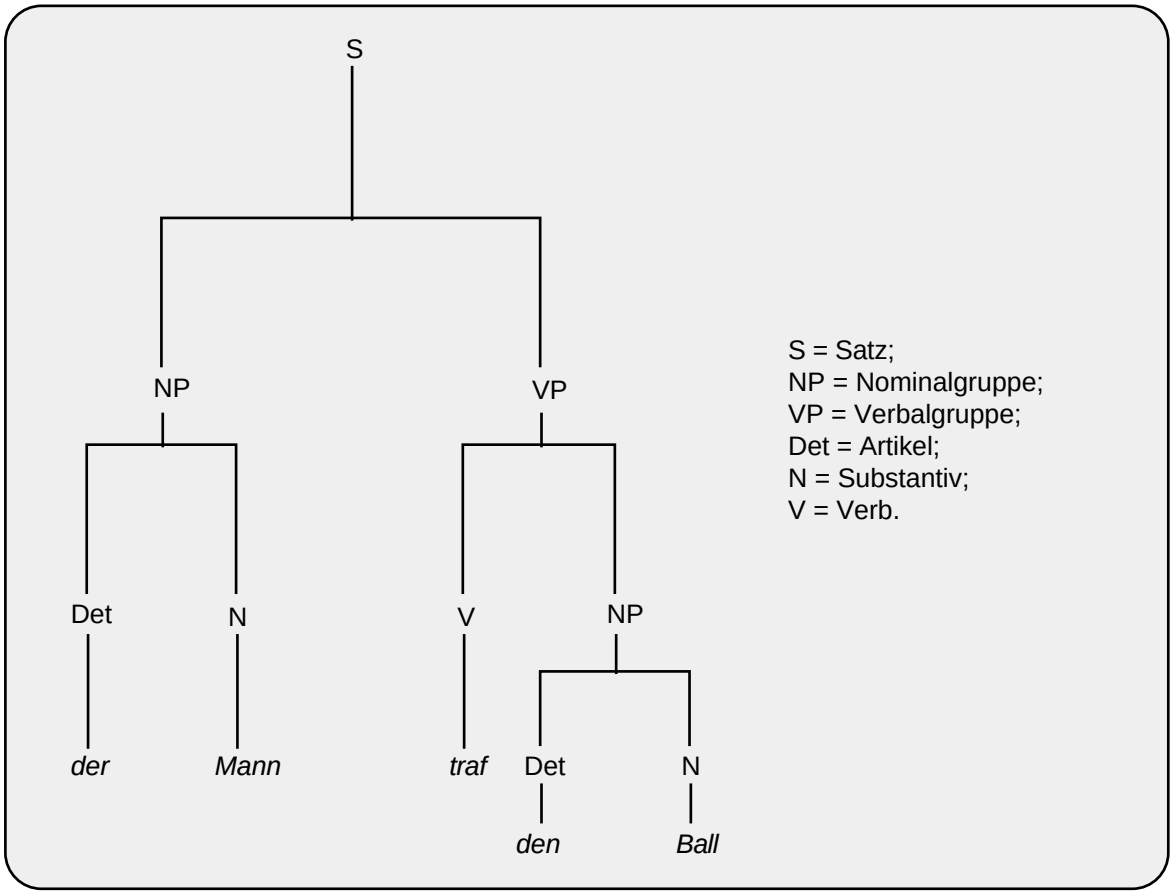


Bild **2.7**: Beispiel einer kontextfreien Phrasenstrukturanalyse (Strukturbaum)

Anhand dieser einfachen Grammatiken lassen sich Nominalgruppen und Verbalgruppen identifizieren, aus denen geeignete Deskriptoren ausgewählt werden können.

Die Nachteile des Verfahrens liegen zum einen darin, daß nicht für jeden Satz ein Strukturbaum erstellt werden kann. Weiterhin können auch Fälle auftreten, bei denen

mehrere Strukturbäume gebildet werden, von denen man nicht weiß, welche semantisch korrekt sind, obwohl sie alle den syntaktischen Regeln entsprechen. Ein dritter Mangel liegt darin, daß semantischen Beziehungen zwischen nicht direkt nebeneinander stehenden Satzkomponenten nicht erkannt werden.

Ein weitaus komplexeres Verfahren arbeitet mit Transformationsgrammatiken. Mit Hilfe von kontextuellen Parametern wird versucht, die mögliche Anzahl der Strukturbäume zu reduzieren.

Der Satzerkennungsvorgang mit Transformationsgrammatiken besteht aus zwei Phasen:

1. Erstellung der Strukturbäume und damit der Oberflächenstruktur mit Hilfe eines Standardparsingsystems.
2. Durch die Umkehrung der Transformation (Transformationsregeln sind kontextsensitive Ersetzungsregeln) wird aus der Oberflächenstruktur die Tiefenstruktur ermittelt. In der Tiefenstruktur spiegelt sich die Bedeutung des Satzes wider.

Ein Problem dieses Verfahrens besteht darin, daß an jeder Stelle der Oberflächenstruktur eine Transformation durchgeführt werden kann und sich die Frage stellt, welche ist die richtige [Allen87], [Sager81].

Ist aber das Vokabular des Diskursgebietes und das syntaktische Muster eingeschränkt, läßt sich eine kanonische Repräsentation der natürlichen Sprache ermitteln, in der die Mehrwortbegriffe mit ihrer Bedeutung erkennbar sind.

2.2. Morphologische Komponente "MORPHIX"

Um beim Indexierungsvorgang die Menge der möglichen Deskriptoren zu verkleinern, wird im klassischen Information Retrieval eine Reduzierung der Textwörter auf ihre Stammform durchgeführt. Dieser Vorgang wird üblicherweise bei englischsprachigen Texten durch eine Abtrennung der Endungen, oftmals mit Hilfe einer Suffixliste, erreicht. Da die deutsche Sprache aber zu einer Sprachklasse gehört, in der es eine freie Wortstellung gibt, verbunden mit mehr grammatikalischer Information der einzelnen flektierten Wörter, ist eine andere Art der Stammformreduzierung notwendig.

Für diese Aufgabe werden morphologische Analyseprogramme eingesetzt, die über linguistisches Wissen der deutschen Sprache verfügen. Die im INFOCLAS-System eingesetzte morphologische Komponente ist das System *MORPHIX* [Finkler86], [Finkler88]. Dieses lexikonbasierte System zur Flexionsanalyse deutscher Sätze wurde im Rahmen eines Fortgeschrittenenpraktikums an der Universität des Saarlandes im Wintersemester 1985/86 in der Programmiersprache LISP entworfen. Das Programm führt eine Lemmatisierung der deutschen Wortformen durch, d.h. eine Zurückführung von flektierten Wortformen auf kanonische Formen (Stammformen) und basiert auf der Verbindung von deklarativem Wissen (Vollformenlexikon) und prozeduralem Wissen (Berechnungsalgorithmen mit einem Grundformenlexikon als deklarativer Wissensbasis). Neben der Fähigkeit der Reduzierung der Wörter auf ihre Wortstämme kann MORPHIX auch zur Generierung von Wortformen in der Sprachübersetzung eingesetzt werden.

Ein Beispiel für eine Analyse durch das MORPHIX-System zeigt Bild 2.8.

Die Eingabe:

```
(morphix ("dieser" "abschnitt" "wird" "immer" "groesser" "!"))
```

liefert das Analyseergebnis:

```
(((("dies"          (WORTART DETERMINATIV)
                    (FLEXION ((MAS ((SG (NOM)) (PL (GEN))))
                    (FEM ((SG (GEN DAT)) (PL (GEN))))
                          (NTR ((PL (GEN))))))))

(("abschneid"       (WORTART VERB) (FLEXION ((IMPERFERKT ((SG (1 3)))))))

("abschnitt"        (WORTART NOMEN)
                    (FLEXION ((MAS ((SG (NOM DAT AKK))))))))

(("werd"            (WORTART VERB) (FLEXION ((PRAESENS ((SG (3))))))))

(("immer"           (WORTART ADVERB)))

(("gross"           (WORTART ADJEKTIV)
                    (FLEXION ((KOM ((PRAEDIKATIV-GEBRAUCHT)))))))
```

```
(("!"          (WORTART INTERPUNKTION)
               (SATZZEICHEN AUSRUFEZEICHEN))))
```

Bild **2.8.**: Analyseergebnis von MORPHIX

Da bei einer Untersuchung der Wörter durch MORPHIX nicht nur die Stammformen berechnet werden, sondern auch die entsprechenden Wortarten, kann die bei der automatischen Indexierung eingesetzte Stoppwortliste der Funktionswörter durch die Analyseergebnisse ersetzt werden. Das heißt, daß nicht die Funktionswörter an sich ausgefiltert werden, sondern viel mehr nur die Wortstämme als Deskriptoren ausgewählt werden, die zu besonderen Wortklassen gehören. Beispiele für Wortklassen sind in diesem Zusammenhang Nomen, Verben, Adjektive usw.

Weil eine Bereitstellung aller deutschen Wörter in MORPHIX schwer realisierbar sein wird, muß eine Eingabemöglichkeit von neuen Wörtern vorhanden sein. Dafür wird eine Wissensakquisitionskomponente zur Verfügung gestellt, die eine halbautomatische Erweiterung der internen Lexika gewährleistet. Mit ihr können Wörter, die nicht in den MORPHIX-Lexika enthalten und damit nicht analysierbar sind, eingetragen werden. Dabei werden grammatikalische Eigenschaften des jeweiligen Wortes abgefragt und intern kodiert (z.B. Genus, Vergangenheitsform, Präfix, Plural usw.). Die vom INFOCLAS-System verwendete Version wurde erweitert um die Wörter der vorliegenden Geschäftsbriefsammlung (100 Exemplare) und den 5000 häufigsten deutschen Wörtern nach einer Statistik von Meier [Meier78].

Bei der Erweiterung und der Benutzung von MORPHIX im Rahmen des INFOCLAS-Systems traten Probleme und Schwierigkeiten auf, die in [Dittrich92] näher erläutert werden. Einige davon werden im folgenden kurz beschrieben.

1) Die Analyse von Wörtern ist nicht immer eindeutig, d.h. es werden mehrere Wortstämme und Wortarten als Ergebnis ausgegeben. Zum Beispiel für das Wort *"führen"* wird folgende Liste (Bild 2.9) ausgegeben:

```
(("fuehr"      (WORTART VERB)
                    (FLEXION ((PRAESENS ((PL (1 3))))
                              (KONJUNKTIV-1 ((PL (1 3))))
                              (INFINITIV)
                              (IMPERATIV (ANREDE)))))
("fahr"        (WORTART VERB)
               (FLEXION ((KONJUNKTIV-2 ((PL (1 3))))))))
```

Bild **2.9.**: Analyseergebnis des Verbs *"führen"*

Dieses Problem ensteht, da in MORPHIX keine kontextuelle Informationen berücksichtigt werden, mit denen eine Einschränkung der Analyseergebnisse möglich wäre. Diese könnte jedoch für die Weiterverarbeitung durch INFOCLAS hilfreich sein zur Vermeidung von Doppel- und Mehrdeutigkeiten.

2) Alle Wörter, die von MORPHIX analysiert werden sollen, müssen in Kleinbuchstaben geschrieben sein. Weiterhin ist notwendig, die Umlaute durch die entsprechenden Doppelbuchstaben zu ersetzen (z.B. ä → ae, ö → oe, ü → ue, ß → ss usw.), was zu einer Verfälschung des Orginaltextes führen und Analysefehler nach sich ziehen kann.

3) Die Analyse von Textkomponenten, wie Abkürzungen (z.B. Dr., usw., etc., G. Müller, u.ä., ...), Bindestrichkomposita (z.B. Hoch- und Tiefbau), Daten (z.B. 27. 5. 1992) und Konstrukten mit "/" (z.B. Bewerber/innen) ist durch MORPHIX nicht möglich. Dies kann zu Informationverlust führen und müßte durch Vorbehandlung eines entsprechenden Experten abgefangen werden.

2.3. Nachrichtentypen (message types)

Mit Verfahren aus dem Gebiet des Information Retrieval werden Hypothesen über den Fokus und der Klasse eines Geschäftsbriefes berechnet. Im weiteren Verlauf des Abschnitts werden diese Klassen, im Rahmen des ALV-Projekts auch Nachrichtentypen (message types) genannt, vorgestellt.

Zuerst soll begründet werden, warum es für die Textanalyse sinnvoll ist eine Einteilung der Geschäftsbriefe in Nachrichtentypen durchzuführen. Ein Grund ist der möglicherweise sehr große Umfang des notwendigen Wissens, welches zur Verfügung stehen muß, um eine Ermittlung der Bedeutung natürlichsprachlicher Texte zu betreiben. Deshalb ist es sinnvoll, den Kontext einzuschränken und damit auch das notwendige Hintergrundwissen. Weiterhin ist es durch Kontexteinschränkung möglich, besondere Hilfsmittel der Analyse zur Verfügung zu stellen, z.B. spezielle Wörterbücher [Hoch&Malburg92] in denen Begriffe des eingeschränkten Kontextes enthalten sind. In unserem speziellen Fall stellen die modellierten Nachrichtentypen einen eingeschränkte Kontext dar [Dittrich92].

Ein weiterer Grund für eine Klassifizierung besteht in einer erwartungsgesteuerten semantischen Analyse von Geschäftsbriefen. Ist der Typ eines Briefes bekannt, so kann ein spezieller Experte für die Textanalyse gestartet werden, welcher eine erwartungsgesteuerte Untersuchung des Briefinhalts ausführt. Dafür werden von dem Experten vordefinierte Ablaufskripts (*Sketchy Scripts* [DeJong82]) über den möglichen Verlauf und Inhalt eines Briefes dieses Nachrichtentyps verwendet.

Prinzipielll ist eine Einteilung der Geschäftsbriefe in vordefinierte Klassen mit Schwierigkeiten behaftet. Eine erste Schwierigkeit tritt in der Festlegung der Art und Anzahl der verwendeten Nachrichtentypen auf. Bei der Art der Klassen stellt sich die Frage, welche Nachrichtentypen werden benutzt, wie sollen sie aussehen und lehnt man sich bei ihrer Definition an internationale Standards an. Bezüglich der Anzahl der Nachrichtentypen muß überlegt werden, wie fein die Struktur der Nachrichtentypen sein soll. Nimmt man nur wenige Nachrichtentypen, so wird die Klassifizierung zwar leichter, aber eine Weiterverarbeitung kann weniger a-priori Wissen einsetzen. Werden die Nachrichtentypen spezifischer bzw. feiner definiert, so wächst ihre Anzahl und eine Klassifizierung wird schwieriger.

Ein weiteres grundlegendes Problem entsteht schon bei einer manuellen Klassifizierung der Geschäftsbriefe. Diese kann nicht in jedem Fall eindeutig durchgeführt werden. Einerseits ist es schwierig für gewisse Fälle überhaupt einen passenden Nachrichtentyp zu finden. Andererseits existieren Geschäftsbriefe, deren Textteile unterschiedlichen Nachrichtentypen angehören. In diesem Fall könnte man dem Brief mehrere Nachrichtentypen zuordnen oder einen dominanten Typ bestimmen, der dann für den gesamten Brief gültig ist. Diese Problematik hängt auch wiederum von der Feinheit der Nachrichtentypdefinition und der Typauswahl ab.

Im konkreten Fall des INFOCLAS-Systems stand eine Geschäftsbriefsammlung von ungefähr 100 Exemplaren zur Verfügung. Aufgrund dieser Sammlung wurde die Festlegung der Nachrichtentypen vorgenommen. Um eine Anbindung an internationale Standards zu erreichen, erfolgte eine Orientierung an dem EDIFACT-Standard [Frank91].

EDIFACT (**E**lectronic **D**ata **I**nterchange **f**or **A**dministration **C**ommerce and **T**ransport) ([ISO8613], [ISO8873], [ISO9735], [Horak85]) ist ein Standard für den Datenaustausch zwischen zwei Kommunikationspartnern, die sich auf elektronischer Ebene über kommerzielle Dokumente (z.B. Geschäftsbriefe) verständigen wollen. Dabei legt der EDIFACT-Standard allgemeine Regeln und Konventionen fest, nach denen dieser Datenaustausch stattfinden muß. Innerhalb des OSI-Referenzmodells (**O**pen **S**ystems **I**nterconnection) der ISO ist das Einsatzgebiet des sich noch in der Entwicklung befindenden Standards in der Anwendungsschicht anzusiedeln (siebte Schicht).

Innerhalb von EDIFACT wurde ebenfalls eine Unterteilung der Dokumente in Nachrichtentypen vorgenommen, an die sich das INFOCLAS-System anlehnt. Weitere Hilfsmittel, die zur Definition der Nachrichtentypen verwendet wurden, waren Modelle aus der Literatur über moderne Korrespondenz. Besonders erwähnt seien Ansätze von Manekeller [Manekeller91] und Grochla [Grochla et al81]. In diesen Modellen werden Geschäftsbriefe sowohl nach inhaltlichen Gesichtspunkten als auch nach vordefinierten Briefteilen für einzelne Briefklassen klassifiziert. Eine detaillierte Beschreibung und ein Vergleich dieser Modelle findet sich in [Stürmer91].

Auf der Basis der drei Grundlagen Geschäftsbriefsammlung, EDIFACT und Modelle aus der Literatur wurde eine Festlegung der vorläufigen Gestalt und Anzahl der Nachrichtentypen für das INFOCLAS-System durchgeführt. Restriktive Vorgabe war jedoch die momentane geringe Anzahl von Geschäftsbriefen, die wiederum nur eine kleine Anzahl von Geschäftsbriefarten beinhaltete. Aus den daraus festgelegten sieben Klassen wurden für das INFOCLAS-System fünf Nachrichtentypen ausgewählt. Diese fünf Nachrichtentypen sind:

i) Anfrage;

ii) Angebot;

iii) Bestellbestätigung;

iv) Bestellung;

v) Werbung.

Für jede der fünf Klassen wurden nachrichtentypspezifische Wortlisten erstellt. Alleinig diese Wortlisten repräsentieren die Nachrichtentypen innerhalb von INFOCLAS. Die Aufstellung der Wortlisten wurde einerseits automatisch, durch Häufigkeitsbestimmung der Wortstämme erreicht und andererseits durch manuelle

Ergänzungen von typspezifischen Wörtern bzw. manuellem Aussortieren von typunspezifischen Wörtern. Durch diese Bearbeitung der Wortlisten wurde eine Einteilung der Listen in Primär-, Sekundär- und Tertiärwörter erreicht. Die Primärwörter sollen dabei die Wörter umfassen, die am spezifischsten für den jeweiligen Nachrichtentyp sind. Die Sekundärwörter, die in der Priorität folgenden Wörter usw. Diese Wortlisten dienen bei den Gewichtsermittlungen des INFOCLAS-Systems, insbesondere im Modul Klassifizierer, als eine der Berechnungsgrundlagen [Dittrich 92].

Eine Erweiterung der Nachrichtentypanzahl ist durch Aufstockung der Geschäftsbriefsammlung durch Briefe aus anderen Klassen oder durch manuelle Untersuchungen möglich. Diese neuen Typen können dann, nach Aufstellung der nachrichtentypspezifischen Wortlisten, dem INFOCLAS-System ohne Schwierigkeiten hinzugefügt werden.

3. Die Architektur des Systems "INFOCLAS"

3.1. Idee und Aufbau

Ziel des ALV-Projektes ist, wie der Name aussagt, ein **A**utomatisches **L**esen und **V**erstehen von Texten. Exemplarisch wird innerhalb des ALV-Projektes die Domäne Geschäftsbriefe untersucht. Ein Teilgebiet bei der Analyse von Geschäftsbriefen ist die Identifikation von Nachrichtentypen (message identification). Um dieses Problem zu lösen, wird einerseits eine Texterkennung (text recognition) und andererseits eine Textanalyse (text analysis) betrieben, wobei diese beiden Phasen eng miteinander in Verbindung stehen.

Das Aufgabengebiet der vorliegenden Diplomarbeit fällt in den Bereich der Textanalyse. Die Aufgabe besteht darin, ausgehend von der Texterkennung gelieferter Hypothesen über Wortkandidaten und zugehöriger logischer Objekte, eine Klassifikation und Fokussierung der Geschäftsbriefe durchzuführen. Logische Objekte werden hierbei von vorhergehenden Modulen mit Hilfe geometrischer Verfahren identifiziert.

Grundlegende Idee der Diplomarbeit ist, Techniken und Verfahren aus dem Gebiet des Information Retrieval [Salton83] zu übernehmen und sie für den Bereich der Textanalyse zu modifizieren. Unterstützt werden diese Methoden dabei von dem Wissen über logische Objekte, Nachrichtentypen und Lexika. Das Verfahren besteht nun darin, die Wörter mit Hilfe einer morphologischen Untersuchung in Deskriptoren und Funktionswörter zu unterteilen und ersteren auf statistischem Wege Gewichte zuzuordnen, welche die Bedeutung der Deskriptoren für den Geschäftsbrief widerspiegeln sollen.

Die Realisierung obiger Überlegungen und Verfahren soll von dem innerhalb dieser Diplomarbeit entwickelten *INFOCLAS-System* erbracht werden. INFOCLAS ist ein Akronym und steht für die englischen Begriffe **in**dexing, **foc**using und **clas**sification. Der Namen impliziert somit die Aufteilung des Systems in drei Basismodule. Diese Module sind:

i) der *Indexierer*;

ii) der *Fokussierer*;

iii) der *Klassifizierer*;

Die Aufgabe des Indexierers ist die Extrahierung und Gewichtung der Deskriptoren. Der Fokussierer verwendet die Ergebnisse des Indexierers als Basis für die Berechnung der Satzteile (Phrasen) oder Sätze im Dokumenttext, welche die Kernaussage des Briefes enthält, da in einem solchen Bereich die wichtigsten und bedeutendsten Wörter vornehmlich auftreten sollten. Der Klassifikator benutzt die gewichteten Deskriptoren zur Festlegung des Nachrichtentyps der Geschäftsbriefe.

Das Zusammenspiel der Module innerhalb des INFOCLAS-Systems wird in der folgenden Darstellung der Systemarchitektur aufgezeigt (Bild 3.1). Dabei werden die benutzten Hilfsmittel und Wissensquellen sowie die Ein- und Ausgabeschnittstellen der Module ebenfalls veranschaulicht.

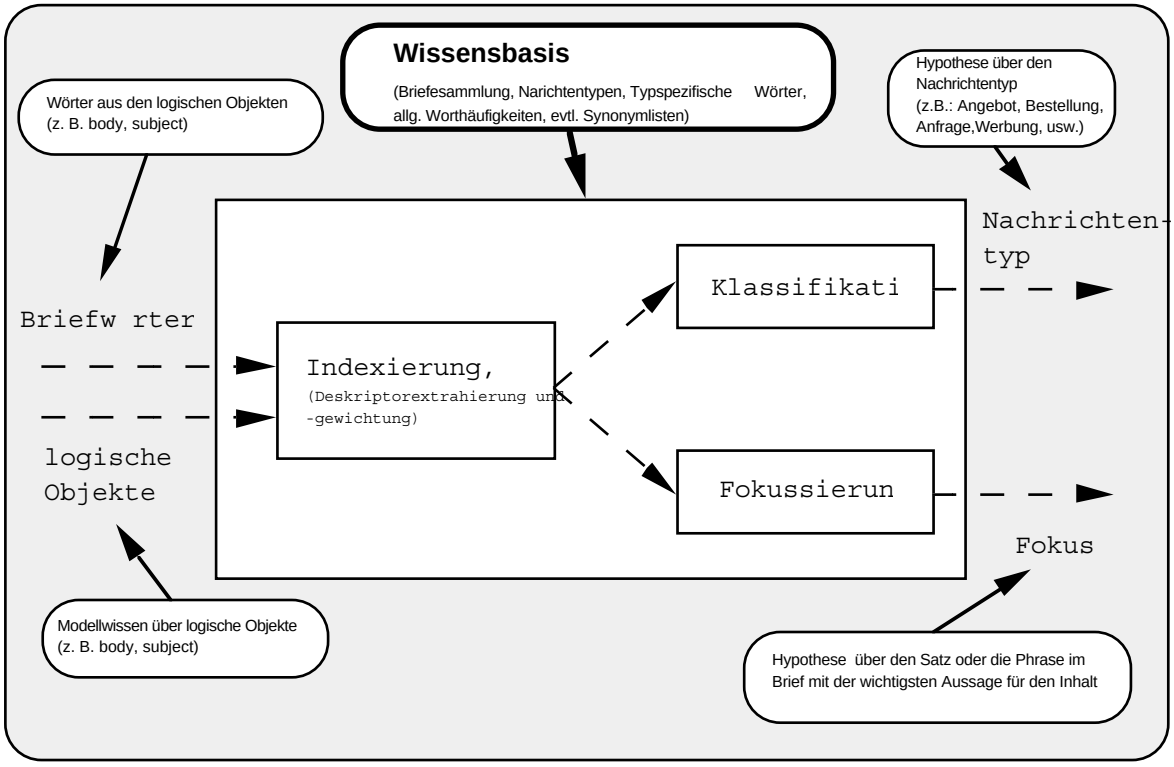


Bild **3.1**: Architektur des INFOCLAS-Systems

Die Bearbeitung eines Briefes beginnt grundsätzlich mit der Auswahl der Deskriptoren aus dem Brieftext sowie der anschließenden Gewichtung. Dazu kann vom Benutzer eine Gewichtsfunktion ausgewählt werden. Benutzt man die Funktionalität des Indexierers, werden die gewichteten Deskriptoren als Ergebnis ausgegeben. Innerhalb der anderen Subsysteme hingegen werden die gewichteten Deskriptoren als Grundlage für weitere Berechnungen benutzt, die dann letztlich zu Fokus- und Klassifikationshypothesen führen. Dem Benutzer wird die Möglichkeit eröffnet, aus diesen zwei Modulen den Indexierer aufzurufen, ohne die Module erst verlassen zu müssen.

Bei der Gewichtung der Deskriptoren mittels vom Indexierer zur Verfügung gestellter Gewichtsfunktionen ist zu beachten, daß die Ergebnisse davon abhängig sind, wieviele Briefe in der Datenbasis gespeichert sind. Einige der Gewichtsfunktionen stellen nämlich Beziehungen zwischen den Wortfrequenzen im Brief und in der Dokumentation her. Die Aussagekraft der Ergebnisse wächst dadurch mit der Anzahl

der Briefe in der Dokumentation, d.h. es sollte eine große Zahl von Briefen vorhanden sein, um "gute" Gewichtungsresultate zu erzielen. Andererseits hat dies zur Folge, daß die Berechnungen dann einem größeren Zeitbedarf unterliegen. Es sind aber in der Datenstruktur der Datenbasis und im Aufbau der internen Suchvorgänge noch Kapazitäten zur Beschleunigung vorhanden, so daß dadurch ein Ausgleich geschaffen werden könnte.

Der Aufbau des INFOCLAS-Systems erlaubt nur die sequentielle Bearbeitung von einem Brief zu einem Zeitpunkt. Sollten mehrere Briefe analysiert werden, müssen sie sequentiell aus der Datenbasis oder neu von Datei geladen werden. Das ist deshalb sinnvoll, da die gewichteten Deskriptoren, der Fokus oder die Klasse immer nur für einen Brief allein bestimmt werden. Einziger Bezug für die Berechnungen ist dabei die Dokumentation (Menge der gespeicherten Briefe) in der Datenbasis und nicht spezielle ausgewählte Briefe.

3.2. Schnittstellenbeschreibung

Die Eingabe des INFOCLAS-Systems ist ein von der Texterkennung analysierter Brief. Da zum Zeitpunkt der Diplomarbeit eine vollständige Texterkennung noch nicht möglich war, mußte die Eingabe für das INFOCLAS-System simuliert werden. Dies hatte zur Folge, daß die Beispielbriefe manuell aufbereitet werden mußten. Ebenso lag die endgültige Form der Schnittstelle zwischen Texterkennung und der darauf aufbauenden Textanalyse noch nicht vor, so daß folgende Vereinbarungen festgelegt wurden.

1) Die Teile des Briefes, die dem INFOCLAS-System zur Analyse übergeben werden, müssen in die Form zweier Lisp-Listen transformiert werden. Die erste Liste enthält den eigentlichen Text des Briefes. Sie ist aus mehreren Unterlisten aufgebaut, welche die logischen Objekte repräsentieren zu denen die übergebenen Textteile gehören. Die logischen Objekte wurden in einer früheren Analysephase (logical labeling) bestimmt. In den Unterlisten befinden sich die Wörter und Satzzeichen in Reihenfolge ihres Auftretens im Brieftext. Die Satzzeichen sind notwendig, um eine Satzstruktur des Textes innerhalb des INFOCLAS-Systems aufbauen zu können. Weiterhin wird davon ausgegangen, daß keine Wildcards in den Wörtern vorkommen, weil nur vollständig erkannte Wörter sinnvoll verarbeitet werden können. Falls doch unvollständige Wörter in logischen Objekten auftreten, werden sie bei der Indexierung nicht von der morphologischen Komponente analysiert und damit als Stoppwörter klassifiziert. Ebenso werden auch keine Wortalternativen oder fehlerhaft erkannte Wörter in dieser Version des Indexierers berücksichtigt.

2) Die zweite Liste der Schnittstellenstruktur der Briefe enthält die Namen der logischen Objekte. Diese wurden durch geometrische Untersuchungen des Brieflayouts (logical labeling) ermittelt. Die Anzahl der Namen muß gleich sein mit der Anzahl der Unterlisten aus der ersten Wortliste. Die Reihenfolge, in der

die Namen der logischen Objekte auftreten, ist entscheidend, da die Namen den Unterlisten paarweise zugeordnet werden. Der erste Name wird der ersten Unterliste zugeordnet, der zweite Name der zweiten Unterliste usw. Damit ist eine genaue Zuordnung der Wortlisten zu den logischen Objekten möglich.

Bild 3.2. zeigt das vollständige Aussehen der Struktur in der ein Brief repräsentiert sein muß, um vom INFOCLAS-System akzeptiert zu werden:

$$((wort_{11}\ wort_{12}\ \ldots\ wort_{1n})$$
$$(wort_{21}\ wort_{22}\ \ldots\ wort_{2m})$$
$$:$$
$$(wort_{s1}\ wort_{s2}\ \ldots\ wort_{st}))$$

$$(sym_1\ sym_2\ \ldots.\ sym_s)$$

Bild **3.2**: Übergabestruktur der Briefe

Bild 3.3 veranschaulicht die Doppellistenstruktur eines Beispielbriefs:

Text:

(("Betr." ":" "Vertrag-Nr." "5-112-4308" "über" "\"" "Anfertigung"
"einer" "Studie" """ "Wissensbank" """ """ """)
("Sehr" "geehrte" "Damen" "und" "Herren" "," "als" "Anlage"
übersenden" "wir" "Ihnen" "den" "Entwurf" "des" "obigen" "Vertrages"
"in" "dreifacher" "Ausfertigung" "mit" "der" "Bitte"
"um" "Unterzeichnung" "und" "Rücksendung" "von" "zwei" "Exemplaren"
"an" "die" "DLR" "," "Hauptabteilung" "Beschaffung" ","
"Linder" "Höhe" "," "5000" "Köln" "90" "."
"Nach" "Gegenzeichnung" "durch" "uns" "erhalten" "Sie"
"ein" "Original" "des" "Vertrages" "für" "Ihre" "Akten" "." "Mit"
"freundlichen" "Grüßen" "i.A." "R" "Derkum")
("Anlagen" ":" "3" "Vertragsexemplare"))

Logische Objekte:

(subject body enclosure)

Bild **3.3**: Doppellistenstruktur eines Beispielbriefs

Jeder Brief, der in obiger Schnittstellenstruktur in einer Datei abgelegt ist, kann vom INFOCLAS-System verarbeitet werden. Wörter und Sätze werden vom System in eine spezielle, interne Datenstruktur umgewandelt (siehe Kapitel 3.3). Diese Transformation dient dazu, einen schnelleren und komfortableren Zugriff auf die für die Briefanalyse relevanten Daten zu gewährleisten. Weiterhin werden statistische Vorberechnungen durchgeführt zur Entlastung späterer Verfahren. Auf diese interne Repräsentation des aktuell zu bearbeitenden Briefes wird von allen drei Subsystemen des INFOCLAS-Systems zugegriffen. Es ist dabei unerheblich in welchem Modul der Brief eingelesen wurde, da von allen dieselbe Datenstruktur aufgebaut wird. Ebenso wird ein aus der Datenbasis eingelesener Brief in die interne Datenstruktur umgewandelt. In der Datenbasis befinden sich neben den schon gespeicherten Briefen und Deskriptoren das Wissen über die modellierten Nachrichtentypen. Weiteres linguistisches Wissen ist in der morphologischen Komponente MORPHIX gespeichert. Hierbei ist zu bemerken, daß die internen Lexika des morphologischen Werkzeugs durch Einlesen neuer Briefe mit Hilfe einer Wissensaquisitionskomponente erweitert werden können. Näheres über MORPHIX wird im Abschnitt 2.2 beschrieben.

Als erstes der drei obigen Module wird im nächsten Kapitel das Kernmodul *Indexierer* vorgestellt. Weiterhin wird auf die interne Repräsentation der Briefe näher eingegangen und die Anwendung von adäquaten Verfahren aus dem Gebiet des Information Retrievals.

3.3. Indexierer

Kern des INFOCLAS-Systems ist das Modul *Indexierer*. Die Methoden des Indexierers stellen eine Voraussetzung für alle weiteren Module des INFOCLAS-Systems dar. Seine Resultate können sowohl als eigenständige Ergebnisse betrachtet werden, als auch von anderen Experten des ALV-Projekts interpretiert und als Grundlage ihrer eigenen Berechnungen herangezogen werden. Bei der Entwicklung des INFOCLAS-Systems und speziell dem Modul Indexierer wurde auf zwei Eigenschaften besonderer Wert gelegt. Zum einen auf ein großes Maß an Parametrisierbarkeit der Verfahren. Zum anderen sollte die Möglichkeit bestehen, die Verfahren entweder automatisch anzustoßen (d.h. von anderen Experten steuerbar) oder sie von einem menschlichen Benutzer mit Hilfe von Menüs zu aktivieren. Die Stellung des Indexierers innerhalb des Systems wird in Bild 3.4 deutlich.

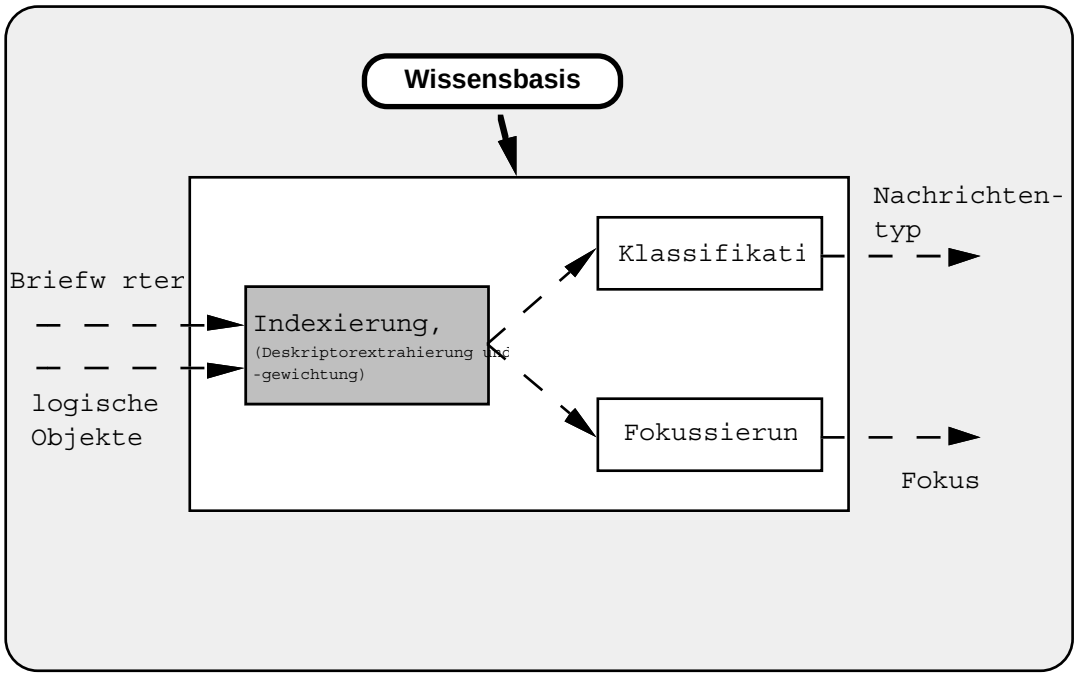


Bild **3.4**: INFOCLAS-System

Mit Hilfe der hohen Parametrisierbarkeit ist es möglich, im Laufe der Entwicklung des ALV-Projekts flexibel auf sich ändernde Anforderungen zu reagieren und damit leicht Modifikationen am System durchzuführen. Weiterhin war ein einfaches Austesten des INFOCLAS-Systems während der Entwicklungs- und Implementierungsphase gegeben. Das war sehr hilfreich, weil die grundlegenden Methoden aus dem Gebiet des Information Retrieval stammen und damit Änderungen zugänglich gemacht werden mußten.

Ein zweigleisiges Konzept (Menüoberfläche/Automatik) ermöglicht eine Nutzung der Funktionalität des INFOCLAS-Systems von Hand, also von einem menschlichen Benutzer, sowie eine leichte Einbindung in die ALV-Programmstruktur (d.h. von anderen Programmen aktivierbar). Positiver Nebeneffekt der Menüoberfläche war ein komfortables Austesten der Systemfunktionalität in der Implementierungs- und Testphase, weil durch die verschiedenartigen Parameter der einzelnen Module zahlreiche Einstellungen getestet werden mußten.

Hauptgegenstand der Berechnung des INFOCLAS-Systems sind die weniger strukturierten Teile von Geschäftsbriefen. Darunter versteht man die Teile eines Briefes, die hauptsächlich natürlichsprachliche Sätze enthalten, insbesondere der Briefrumpf. Ein weiterer Unterschied gegenüber stark strukturierten Briefteilen ist eine größere Anzahl von Wörtern. Unter strukturierten Elementen kann man sich logische Objekte vorstellen wie Empfängeradresse, Senderadresse, Datum, Firmenlogo, u.ä. Sie enthalten keine vollständigen Sätze und in der Regel wesentlich weniger Wörter als die restlichen Teile eines Geschäftsbriefes. Zudem versprechen hier andere, z.B. syntatisch orientierte Analyseansätze, die nicht auf gewichteten Deskriptoren basieren, mehr Erfolg.

Anwendungsgebiet des Indexierers ist damit im wesentlichen das logische Objekt *Briefrumpf*, da hier natürlichsprachliche Sätze und der Hauptanteil der Wörter des Briefes zu finden sind. Weiterhin besteht die Möglichkeit, andere logische Objekte wie Betreffteil, Postscript, Anlage, Verteiler u.ä. bei der Untersuchung durch den Indexierer zu berücksichtigen. Prinzipiell können auch stark strukturierte Teile von diesem Modul bearbeitet werden. Es ist aber zweifelhaft, ob dies sinnvoll ist, gerade im Hinblick auf eine spätere Verwendung der Indexierergebnisse innerhalb des INFOCLAS-Systems (Fokus- und Nachrichtentypbestimmung).

Nachfolgend wird die Arbeitsweise und Funktionalität des Indexierers beschrieben. Die oben genannten Briefelemente, die vom Indexierer analysiert werden sollen, müssen hierfür von der Texterkennung aufbereitet und in dem INFOCLAS-spezifischen Listenformat (vgl. Kapitel 3.2.) zur Verfügung gestellt werden. Um eine bessere Analyse der Briefwörter durch die morphologische Komponente zu erhalten, sollte eine zusätzliche Aufbereitung nach den in Kapitel 2.2. aufgeführten Sonderfällen erfolgen. Dazu gehören z.B. Abkürzungen, Bindestrichkomposita, Datum, Zahlen, Konstrukte wie "er/sie" usw. Durch eine Auflösung dieser Sonderfälle bzw. Trennung der Wörter ist eine größere Informationserschließung aus dem Brieftext möglich.

In den folgenden Abschnitten wird die Beschreibung der Funktionsweise des Indexierers anhand eines konkreten Beispiels unterstützt. Als Untersuchungsobjekt dient dazu der Brief aus Kapitel 3.2. Der Brief wurde zuvor von der Texterkennung analysiert und liegt im Doppellistenformat vor. Er beinhaltet somit Informationen über die Briefwörter, die Satzstruktur und die logischen Objekte. Beim Start des Indexieres, wie auch bei den anderen zwei Modulen des INFOCLAS-Systems, wird beim Einlesen eines Briefes eine Instanz der Klasse *letter* erzeugt. Diese Instantiierung wird unabhängig davon ausgeführt, ob der Brief von Datei neu eingelesen oder aus der

Wissensbasis geladen wurde. Bei diesem Vorgang werden die einzelnen Slots der Instanz mit konkreten Werten aus dem Brief gefüllt. Eine Instanz der Klasse *letter* besitzt acht Slots, die im Bild 3.5 schematisch gezeigt und im folgenden einzeln vorgestellt werden.

Slot	Sloteintrag
name	Briefname
number	Briefnummer
log-object	Liste der logischen Objekte
sents	Liste der Anzahl der vorhandenen Sätze
words	Liste der Anzahl der vorhandenen Wörter
wordlist-all	Liste der Briefwörter in Satz- und logische Objektstruktur
wordlist-morph	Liste wie in wordlist-all, Wörter morphologisch untersucht und Satzzeichen entfernt
wordlist-red	Deskriptorliste

BILD **3.5**: Interne Datenstruktur

Die ersten beiden Slots dienen der Identifikation des Briefes. Der Benutzer kann dem aktuellen Brief einen eindeutigen Namen zuordnen, der im Slot *name* abgelegt wird. Zur internen Identifikation erhält der Brief zusätzlich eine eindeutige Briefnummer vom System zugeordnet; sie wird im Slot *number* vermerkt.

Die restlichen sechs Instanzeneinträge beinhalten Wort- und Statistikinformationen aus dem Briefinhalt. Im Slot *log-object* sind die Anzahl und die Namen der logischen Objekte eingetragen, die dem INFOCLAS-System zur Analyse übergeben wurden. In den Einträgen *sents* und *words* werden jeweils die Anzahl der Wörter und Sätze angegeben sowie die Häufigkeit ihres Auftretens in den einzelnen logischen Objekten. Der sechste Slot *wordlist-all* beinhaltet die vollständige Wort- und Satzinformation des Briefes. Das bedeutet, daß die Wörter und Satzzeichen in der von der Texterkennung aufbereiteten Form in Satz- und Objektlisten vorliegen. Wegen der morphologischen Komponente MORPHIX mußten zwei Änderungen an den Wörtern während des Einlesevorgangs vorgenommen werden. Erstens wurden alle Großbuchstaben in Kleinbuchstaben umgewandelt und zweitens die Umlaute durch gleichbedeutende Doppelbuchstaben, z.B. ä → ae, ö → oe, ü → ue und ß → ss, ersetzt. Eine sinnvolle Erweiterung von MORPHIX wäre in diesem Zusammenhang, wenn auch Wörter mit Großbuchstaben und Umlauten analysiert werden könnten. Der Aufbau des Sloteintrages ist in diesem Falle eine Liste mit Unterlisten für jedes logische Objekt.

Diese wiederum enthalten ebenfalls Unterlisten für jeden Satz, der in einem logischen Objekt enthalten ist. In diesen Satzlisten sind die Wörter und Satzzeichen durch Strings repräsentiert.

Auf die Satzstruktur wird mit Hilfe folgender Informationen geschlossen. Die Satzzeichenstrings ".", "!", "?" sowie Anfang und Ende der logischen Objekte bilden die Grundlagen zur Erkennung von Satzanfang bzw. Satzende. Weitere Satzzeichen, wie Komma, Semikolon u. ä., werden in der aktuellen Version des Systems nicht berücksichtigt. Diese Satzzeichen könnten aber in einer Erweiterung zur Erkennung von Satzteilen, Nebensätzen und Phrasen verwendet werden.

Für die Ermittlung des siebten Sloteintrages *wordlist-morph* wird die morphologische Komponente MORPHIX verwendet. Die Listenstruktur des Slots ist die gleiche wie im vorhergehenden Slot, nur wurden die Wörter in den Satzlisten morphologisch analysiert. Weiterhin sind sämtliche Satzzeichen entfernt. Durch die morphologische Analyse ist es nun möglich, die Wörter durch die zugehörige Stammform zu ersetzen. Für die Ersetzung der Wörter wurde dabei die erste Alternative der MORPHIX-Analyse ausgewählt, falls es Doppel- und Mehrdeutigkeiten für das Wort gab. Eine Ausdehnung wäre hier möglich, indem alle alternativen Wortstämme für ein Wort berücksichtigt würden. Eine andere Lösung für das Problem wäre die Erweiterung der morphologischen Komponente zur Erkennung einer eindeutigen Stammform für jedes Wort. Dafür wäre aber notwendig, daß kontextuelle Informationen bei der Analyse herangezogen werden müßte.

Der achte und letzte Slot *wordlist-red* umfaßt die Liste der Deskriptoren, die für den aktuellen Brief extrahiert wurden. Um diese Deskriptoren zu ermitteln, wird in klassischen Information Retrieval Systemen eine Stoppwortliste verwendet. Die Liste enthält Funktionswörter aus Wortgruppen wie Konjunktionen, Pronomen, Partikel usw. Lediglich nicht in der Stoppwortliste vorkommende Wörter sind Deskriptorkandidaten und werden weiter untersucht (näheres dazu in Kapitel 2.1.).

Im Falle des INFOCLAS-Systems wird die Reduzierung der Stoppwörter durch die Analyseergebnisse der morphologischen Komponente gesteuert. Hierfür wird die bei der Wortuntersuchung ermittelte Wortkategorie verwendet. Kriterium für die Auswahl der Deskriptoren ist die Zugehörigkeit zu den Wortarten Nomen, Verb, Adjektiv und attributiv-gebrauchtes Partizip-Perfekt. Eine sinnvolle Erweiterung des Systems wäre die Berücksichtigung von Wörtern wie Eigennamen, Produktnamen, Ortsnamen u. ä. Sie müßten entweder als Nomen in das MORPHIX-Lexikon eingetragen werden oder im Vorfeld von einem Experten erkannt und als eigene Wortart klassifiziert werden. Ebenso wäre die Hinzunahme der Wörter wichtig, die nicht in Morphixlexika vertreten sind. Dieses wäre aber nur dann sinnvoll, wenn man davon ausgehen könnte, daß alle oder zumindest der größte Teil der deutschen Funktionswörter erfaßt wurden. In diesem Fall wäre nämlich ein neues, unbekanntes Wort ein potentieller Deskriptor und würde zumindest in der vorliegenden Form zur Indexierung verwendet werden.

Der Inhalt des Sloteintrages *wordlist-red* ist eine alphabetisch geordnete Liste mit einer Liste für jeden Deskriptor. Diese Deskriptorlisten enthalten die Stammform des Wortes, die Wortart und die Frequenz des Auftretens der Stammform im aktuellen Brief.

Eine Instanz des aktuellen Briefes mit obig beschriebenen acht Slots bildet die Grundlage für alle weiteren Berechnungen des Indexierers und dem gesamten INFOCLAS-System. In Bild 3.6 wird der Beispielbrief aus Kapitel 3.2. in der internen Repräsentation gezeigt.

Slot	Sloteintrag
name	Brief48
number	27
log-object	(3 (SUBJECT BODY ENCLOSURE))
sents	(5 ((SUBJECT 1) (BODY 3) (ENCLOSURE 1)))
words	(81 ((SUBJECT 14) (BODY 63) (ENCLOSURE 4)))
wordlist-all	(((1 ("Betr." ":" "Vertrag-Nr." "5-112-4308" "ueber" "\"" "Anfertigung" "einer" "Studie" "" "Wissensbank" "" "" ""))) ((1 ("Sehr" "geehrte" "Damen" "und" "Herren" "," "als" "Anlage" "uebersenden" "wir" "Ihnen" "den" "Entwurf" "des" "obigen" "Vertrages" "in" "dreifacher" "Ausfertigung" "mit" "der" "Bitte" "um" "Unterzeichnung" "und" "Ruecksendung" "von" "zwei" "Exemplaren" "an" "die" "DLR" "," "Hauptabteilung" "Beschaffung" "," "Linder" "Hoehe" "," "5000" "Koeln" "90" ".")) (2 ("Nach" "Gegenzeichnung" "durch" "uns" "erhalten" "Sie" "ein" "Original" "des" "Vertrages" "fuer" "Ihre" "Akten" ".")) (3 ("Mit" "freundlichen" "Gruessen" "i.A." "R" "Derkum"))) ((1 ("Anlagen" ":" "3" "Vertragsexemplare"))))
wordlist-morph	(((1("betr" "vertrag-nr" "5-112-4308" "ueber" "\"" "anfertigung" "ein" "studie" "wissensbank"))) ((1 ("sehr" "ehr" "dame" "und" "herr" "als" "anlage" "uebersend" "wir" "sie" "d-" "entwurf" "d-" "obig" "vertrag" "in" "dreifach" "ausfertigung" "mit" "d-" "bitt" "um" "unterzeichnung" "und" "ruecksendung" "von" "zwei" "exemplar" "an" "d-" "dlr" "hauptabteilung" "beschaffung" "linder" "hoehe" "5000" "koeln" "90")) (2 ("nach" "gegenzeichnung" "durch" "wir" "erhalt" "sie" "ein" "original" "d-" "vertrag" "fuer" "ihr" "akt")) (3 ("mit" "freundlich" "gruss" "i" "r" "derkum"))) ((1 ("anlage" "3" "vertragsexemplar"))))

```
wordlist-red    (("akt" MORPHIX:NOMEN 1)
                 ("anfertigung" MORPHIX:NOMEN 1)
                 ("anlage" MORPHIX:NOMEN 2)
                 ("ausfertigung" MORPHIX:NOMEN 1)
                 ("beschaffung" MORPHIX:NOMEN 1)
                 ("bitt" MORPHIX:VERB 1)
                 ("dame" MORPHIX:NOMEN 1)
                 ("dreifach" MORPHIX:ADJEKTIV 1)
                 ("ehr" MORPHIX:ATTRIBUTIV-GEBRAUCHTES 1)
                 ("entwurf" MORPHIX:NOMEN 1)
                 ("erhalt" MORPHIX:VERB 1)
                 ("exemplar" MORPHIX:NOMEN 1)
                 ("freundlich" MORPHIX:ADJEKTIV 1)
                 ("gegenzeichnung" MORPHIX:NOMEN 1)
                 ("gruss" MORPHIX:NOMEN 1)
                 ("hauptabteilung" MORPHIX:NOMEN 1)
                 ("herr" MORPHIX:NOMEN 1)
                 ("hoehe" MORPHIX:NOMEN 1)
                 ("obig" MORPHIX:ADJEKTIV 1)
                 ("original" MORPHIX:NOMEN 1)
                 ("ruecksendung" MORPHIX:NOMEN 1)
                 ("studie" MORPHIX:NOMEN 1)
                 ("uebersend" MORPHIX:VERB 1)
                 ("unterzeichnung" MORPHIX:NOMEN 1)
                 ("vertrag" MORPHIX:NOMEN 2)
                 ("vertragsexemplar" MORPHIX:NOMEN 1)
                 ("wissensbank" MORPHIX:NOMEN 1))
```

Bild **3.6**: Interne Repräsentation des Beispielbriefes

Auf die einzelnen Slots eines Briefes wird im Laufe der Indexierung, Fokussierung und Klassifizierung zugegriffen, um jeweils die notwendige Information über den Brief zu erhalten.

Im folgenden soll nun der Bearbeitungsabschnitt der Deskriptorgewichtung etwas genauer beleuchtet und vorgestellt werden. Das Berechnen der Deskriptorgewichte nach ihrer Bedeutung ist die Hauptaufgabe des Indexierers. Diese Gewichte können entweder einzelnen Briefdeskriptoren oder allen Briefdeskriptoren zugewiesen werden. Zur Verfügung stehen dafür verschiedene Arten von Gewichtsfunktionen und spezielle Erweiterungen. Die Gewichtsfunktionen basieren auf Konzepten aus dem Gebiet des Information Retrieval (Kapitel 2.1.3.) und umfassen:

a) *relative und absolute Frequenz innerhalb der Dokumentation;*

b) *Inverse Dokumenthäufigkeit;*

c) *Informationswert;*

d) *Diskriminanzwert.*

Die Gewichtsfunktion *Diskriminanzwert* liegt nur in eingeschränkt implementierter Form vor und ist deshalb noch nicht in das INFOCLAS-System integriert worden. Implementiert ist eine Gewichtung eines einzelnen Deskriptors unter Zuhilfenahme von fünf existierenden Ähnlichkeitsfunktionen.

Unter einer *Dokumentation* versteht man in diesem Zusammenhang alle Briefe, welche sich in der Datenbasis befinden und als Berechnungsgrundlage für die Gewichtsfunktionen herangezogen werden können.

Mit möglichen Erweiterungen werden einzelne Deskriptoren bestimmt, deren Gewichte mit vorher festgelegten Multiplikatoren verrechnet werden, zur Erhöhung bzw. Verminderung ihres Wertes. Eine Erweiterung ist der Abgleich der Deskriptoren mit den *Primär-* und *Sekundärwörtern* der einzelnen, modellierten Nachrichtentypen. Wird eine Übereinstimmung eines Deskriptors mit einem Eintrag in einer der Wortlisten festgestellt, so findet eine Erhöhung des Deskriptorgewichts durch Verrechnung mit dem Multiplikator, der der Wortliste zugeordnet ist (näheres in Kapitel 2.3 und 3.5).

Eine zusätzliche Erweiterung könnte über die statistischen Angaben gewonnen werden, die durch die Wortstatistik von Meier [Meier78] gegeben sind. Meier gibt in seiner Statistik die 8000 häufigsten, deutschen Wörter an. Die Idee für die Erweiterung besteht nun darin, einem Deskriptor, der in der Meierstatistik als ein besonders häufig im Deutschen verwendetes Wort angegeben ist, aber innerhalb eines Dokumentes nur selten vorkommt, mit einen verminderten Wert zu gewichten. Ein Deskriptor, der im Dokument vermehrt vorkommt, aber in der Meierstatistik nicht erscheint, sollte höher eingestuft werden. Dabei sollte aber bedacht werden, daß ein Wort, das in einem Brief nur ein- oder zweimal vertreten ist, schon zu den häufigeren Wörtern zählt und auch eine große Bedeutung für den Brief besitzen kann.

Nach Auswahl einer Gewichtsfunktion oder einer optionalen Erweiterung werden vom Indexierer die Deskriptorgewichte berechnet und paarweise mit den Deskriptoren (Stammform eines Wortes) ausgegeben.

Somit ergibt sich für den Indexierer die in Bild 3.7 dargestellte Bearbeitungsabfolge.

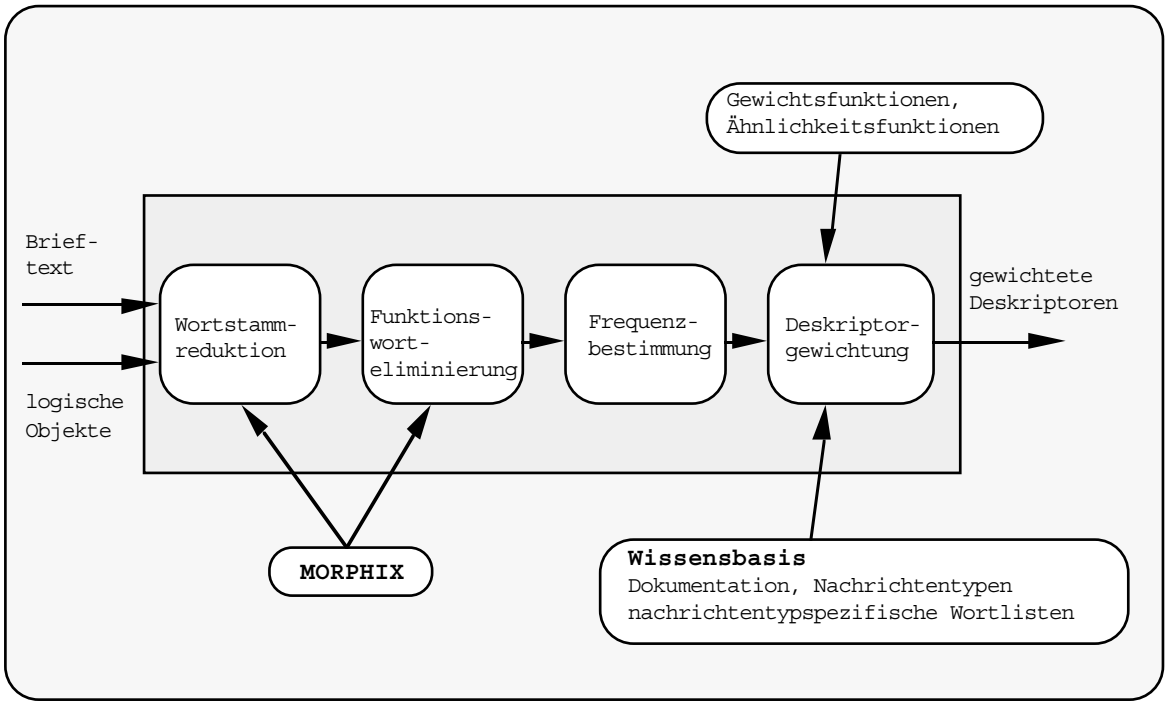


Bild **3.7**: Bearbeitungsabfolge des Indexierers

An dieser Stelle sollte ausdrücklich darauf hingewiesen werden, daß die aus dem Information Retrieval benutzten Konzepte einer anderen Intention unterliegen als jener, die der Dokumentenanalyse zugrunde liegt. Im Information Retrieval sollen Deskriptoren und deren Gewichte ermittelt werden, die spezifisch für genau ein Dokument sind. Die Deskriptoren dienen zur Abgrenzung des einzelnen Dokuments von anderen Dokumenten, um eine möglichst exakte Antwort auf eine Benutzeranfrage (query) geben zu können. Im Fall der Dokumentanalyse hingegen sollen die Deskriptoren den Inhalt des Dokuments widerspiegeln und die Gewichte das Maß ihrer Charakteristik. Dabei ist die Bedeutung der Deskriptoren für andere Dokumente von untergeordneter Bedeutung. Diese Bedeutung wird erst auf der Ebene der Nachrichtentypen und deren spezifischen Wortlisten relevant.

Eine weitere Möglichkeit der Deskriptorgewichtung wäre die Vermischung von ausgewählten Gewichtsfunktionen, die in einem vom Benutzer zu bestimmenden Verhältnis verrechnet würden. Voraussetzung dafür ist eine Angleichung der Wertebereiche der Gewichtsfunktionen. Ein Problem, welches durch diese Art der Gewichtung auftreten könnte, wäre ein gegenseitiges Aufheben der Gewichtungen und damit ein Verlust der Ausagekraft des Ergebnisses.

Zur Veranschaulichung der Bearbeitung von Geschäftsbriefen durch den Indexierer werden in Bild 3.8 die Ergebnisse eines Indexiervorgangs für den Beispielbrief gezeigt. Verwendete Grundlage ist eine Dokumentation mit dem Umfang von 20 Briefen. Zur Gewichtung der Deskriptoren wird die Funktion *Inverse Dokumenthäufigkeit* benutzt.

akt	5.39
anfertigung	5.39
anlage	3.87
ausfertigung	5.39
beschaffung	5.39
bitt	1.69
dame	4.39
dreifach	5.39
ehr	1.07
entwurf	5.39
erhalt	2.39
exemplar	3.39
freundlich	1.07
gegenzeichnung	5.39
gruss	1.07
hauptabteilung	5.39
herr	1.14
hoehe	5.39
obig	5.39
original	5.39
ruecksendung	4.39
studie	4.39
uebersend	4.39
unterzeichnung	5.39
vertrag	10.78
vertragsexemplar	5.39
wissensbank	5.39

Bild **3.8:** Ergebnis des Indexiervorgangs für den Beispielbrief

Das nächste Kapitel stellt das zweite Modul des INFOCLAS-Systems vor, den Fokussierer, der die Ergebnisse des Indexierers als Basis für seine Berechnungen benötigt. Es werden Ideen, Arbeitsweise und Parameter der benutzten Konzepte vorgestellt.

3.4. Fokussierer

Das zweite wichtige Modul des INFOCLAS-Systems ist der *Fokussierer*. Der Fokussierer liefert als Analyseergebnis eine Hypothese über den Satz oder die Sätze, welche die Kernaussage des Briefes enthalten. Somit soll der Fokussierer denjenigen Teil des Briefes extrahieren, der die vermutlich wichtigsten Informationen über die Intention des Briefschreibers enthält. Hypothesen über einen möglichen Fokus können im weiteren von anderen Experten des ALV-Systems (z.B. Insel-Parser, Feature-basierte Parser etc.) als Einstiegspunkt für ihre Berechnungen benutzt werden. Bild 3.8 zeigt die Stellung des Fokussierers im INFOCLAS-System.

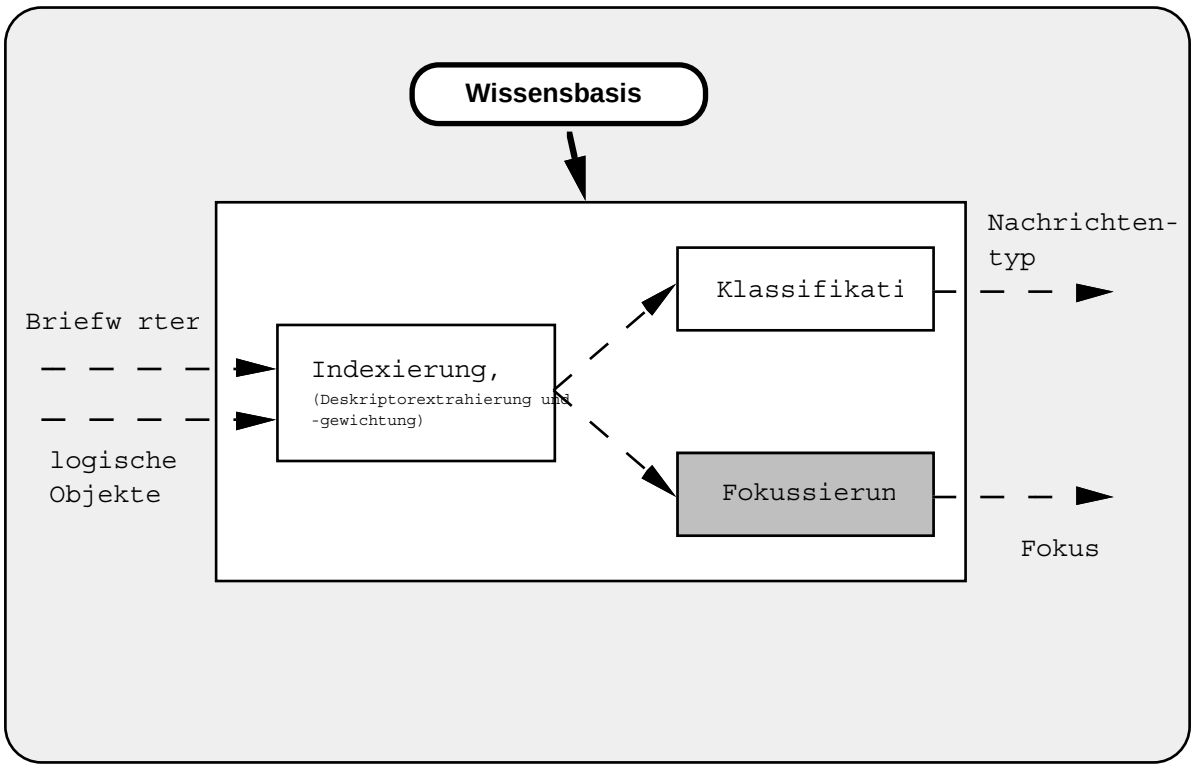


Bild **3.9**: INFOCLAS-System

Grundlagen des Fokussierungsvorgangs sind der zu bearbeitende Brief und die mit Hilfe des Indexierers extrahierten Deskriptoren. Voraussetzungen an die Eingabeschnittstelle sind die gleichen wie im Modul Indexierer, d.h. der aktuelle Brief muß in einer Doppellistenform vorliegen und wird beim Einlesevorgang in die interne Repräsentation transformiert (vgl. Abschnitt 3.3.). Auch hier ist es sinnvoll, eine Fokussierung der weniger strukturierten Briefteile, welche natürlichsprachliche Sätze enthalten, durchzuführen.

In der aktuellen Version des Fokussierers wurde als kleinste Größe für den Fokus der *Satz* gewählt. Dieses kann aber im Bedarfsfall auch dahingehend geändert werden, daß Phrasen oder Deskriptorenbereiche bestimmter Größe als Fokus definiert werden können. Dies könnte durch ein Sliding-window Verfahren realisiert werden.

Der Fokus kann vom Benutzer des Systems satzweise vergrößert werden, falls das ausgewählte logische Objekt oder die Größe des Briefes es zuläßt. Diese Einschränkung auf Satzgröße liegt nahe, da in der internen Repräsentation die Unterteilung der logischen Objekte in Sätze erfolgt. Dazu wurde ein einfaches Verfahren benutzt, welches als Anforderung an die Texterkennung eine korrekte Ermittlung der Satz- und Interpunktionszeichen stellt. Diese müssen dann durch Strings repräsentiert vorliegen. Der Anfang und das Ende eines Satzes wird durch das Verfahren folgendermaßen definiert:

i) Satzanfang

   Das erste Wort des Satzes steht am Anfang des Briefes oder eines logischen Objekts bzw. der Satz beginnt mit dem ersten Wort nach dem Vorgängersatz.

ii) Satzende

   Das letzte Wort des Satzes steht am Ende des Briefes oder eines logischen Objekts bzw. der Satz endet mit einem Punkt, einem Ausrufezeichen oder einem Fragezeichen.

Insbesondere legt obige Definition fest, daß es sich bei einer Gruppe von Wörtern ohne Punkt, Ausrufezeichen oder Fragezeichen um einen Satz handelt, z.B. bei einem einzeiligen Betreffteil oder einer Anlage. Voraussetzung für eine korrekte Satzerkennung ist, daß nur solche Punkte als Strings im Brieftext repräsentiert werden, deren Funktion die Kennzeichnung des Satzendes ist. Andere Punkte müssen an ihre Wörter bzw. Buchstaben gebunden sein, z.B. "bzw.", "usw.", "1.1.1992", "Dr.", "G." "Müller" und ähnliches. Eine mögliche Erweiterung des Fokussierers könnte die Verkleinerung der Mindestgröße für den Fokus sein. Hierfür müßte nur die Abfrage um die Erkennung von Kommas, Semikolons und anderer Satzzeichen ausgedehnt werden oder eine Erweiterung zur Festlegung von Deskriptorbereichen (sliding-window). Damit wäre eine Bearbeitung von Satzteilen, Nebensätzen und Phrasen möglich.

Grundidee der Fokussierung ist die Berechnung von Gewichten für die kleinste Einheit, die der Fokus einnehmen kann. Dieses Satzgewicht wird dadurch ermittelt, daß jedem Deskriptor eines Satzes ein Gewicht zugeordnet wird, die dann aufsummiert werden. Da nun die vorliegenden Sätze i.a. nicht alle gleich lang sind bzw. die gleiche Anzahl von Deskriptoren besitzen, ist eine Normierung der Satzgewichte nötig. Würde man das einfache Satzgewicht durch Aufsummierung verwenden, bekämen lange Sätze im Durchschnitt höhere Werte als kurze . Es ist aber nicht immer so, daß lange Sätze mehr Information enthalten müssen als kurze Sätze. Zur Normierung des einfachen Satzgewichtes werden dem Benutzer deshalb zwei Möglichkeiten angeboten. Die erste

besteht in der Division des Satzgewichtes durch die Anzahl aller Wörter des Satzes oder zweitens die Division durch die Anzahl der Satzdeskriptoren.

Alle weiteren Berechnungen des Fokussierers umfassen nun die größenmäßige Auflistung der Sätze in Reihenfolge ihrer Satzgewichte. Das heißt, der Satz mit dem größten Gewicht hat die größte Wahrscheinlichkeit für die Fokushypothese. Bei einer Fokusberechnung mit einer größeren Satzeinheit werden Satzgruppen gebildet und Satzgruppengewichte berechnet. Die Ermittelung der Gewichte für die Satzgruppen unterliegt der Konvention, daß nur aufeinanderfolgende Sätze eine Satzgruppe und damit einen Fokus bilden können. Es erscheint nicht sinnvoll, Sätze aus verschiedenen Teilen des Briefes zusammen einen Fokus bilden zu lassen, gerade hinsichtlich einer späteren Verwendung durch andere Experten innerhalb des ALV-Systems (z.B. Insel-Parser).

Die Satzgruppengewichtung wird in zwei Stufen durchgeführt:

In der ersten Stufe werden die Gewichte der einzelnen Sätze durch Addition berechnet. Danach wird in einer zweiten Berechnungsstufe ein Fenster in der Größe des zu ermittelnden Fokus über die Satzgewichte geführt. Die sich innerhalb des Fensterausschnitts befindlichen Satzgewichte werden summiert und bilden das Satzgruppengewicht. Durch dieses Verfahren wird gesichert, daß lediglich direkt aufeinanderfolgende Sätze eine Gruppe bilden können. In diesem Zusammenhang ist zu beachten, daß bei der Auswahl von mehreren logischen Objekten zur Fokussierung diese zu einem Objekt vereinigt werden und die ausgegebenen Satznummern sich auf dieses künstliche logische Objekt beziehen.

Aus den beschriebenen Verfahren der Fokussierung ergeben sich mehrere Parameter, die vom Benutzer bestimmt werden müssen. Diese Parameter sind:

i) das logische Objekt (Basis der Berechnung);

ii) die Gewichtsfunktion (Deskriptorgewichtung);

iii) die Fokusgröße (Anzahl der Sätze);

iv) der Divisor (Normierung der Satzgewichte).

Der erste Parameter ist die Angabe des logischen Objektes, aus dem der Fokus bestimmt werden soll. Zur Auswahl stehen prinzipiell alle an das INFOCLAS-System übergebenen, logischen Objekte des Briefes. Weiterhin kann als Berechnungsgrundlage der gesamte Brief gewählt werden, also eine Zusammenfassung der vorhandernen logischen Objekte. Eine zusätzliche Möglichkeit ist die gezielte Auswahl von bestimmten logischen Objekten durch den Benutzer, der dadurch eine Verschmelzung dieser einzelnen logischen Objekte zu einem neuen künstlichen Objekt durchführt. Hiermit besteht für den Benutzer die Option, sich auf speziell für ihn wichtige und interessante Teile des Briefe zu beschränken.

Mit dem zweiten Fokussierungsparameter wird die für alle Berechnungen zugrundeliegende Gewichtsfunktion festgelegt. Über diesen Parameter wird vom Fokussierer aus das Modul Indexierer angestoßen, welches als Ergebnis die gewichteten Deskriptoren liefert. Diese werden dann zur Berechnung der Satzgewichte und Satzgruppengewichte herangezogen. Zu diesen Gewichtungsfunktionen sind wie im Indexierer das Aktivieren von Erweiterungen möglich. Zur Auswahl steht die schon in Kapitel 3.3. vorgestellte Erhöhung der Deskriptorgewichte durch Abgleich mit den Primär- und Sekundärwortlisten der Nachrichtentypen. Als weitere Option wird eine Vorklassifizierung angeboten. Hierbei wird vom Modul Fokussierer das Modul Klassifizierer angestoßen.

Als Ergebnis der Vorklassifizierung, die mit Defaultparametern durchgeführt wird, ergibt sich eine Hypothese über den Nachrichtentyp des Briefes. Diese Information wird verwendet, um den Deskriptoren, die im ausgewählten logischen Objekt und in den Primär- und Sekundärwortlisten des vorklassifizierten Nachrichtentyps auftreten, ein höheres Gewicht zuzuordnen. Die Realisierung wird wiederum durch Multiplikatoren vorgenommen, die vom Benutzer einstellbar sind. Für jede Wortliste wird dabei ein Multiplikator angeboten. Die Idee der Vorklassifizierung ist die, daß im Fokus eines Briefes mit hoher Wahrscheinlichkeit Deskriptoren auftreten, die charakteristisch für den entsprechenden Nachrichtentyp des Briefes sind. Die höhere Gewichtung dieser Deskriptoren erscheint dadurch sinnvoll und unterstützt das Erkennen des Fokus. Eine Gefahr besteht aber in der Fehlerhaftigkeit der Vorklassifizierung, da bei einer falschen Klassifizierung der Fehler sich auch in der Fokussierung niederschlägt. Eine Möglichkeit der Verbesserung wäre eine Angabe einer Prioritätsliste der Nachrichtentypen von der Vorklassifizierung, um mehrere Fokussierungsberechnungen mit verschiedenen Nachrichtentypen durchzuführen. Die zugrundeliegenden Verfahren und Methoden des Moduls Klassifizierer werden im anschließenden Abschnitt 3.5. näher beschrieben.

Der dritte Parameter enthält die Anzahl der Sätze, die der Fokus umfassen soll. Untere Grenze ist ein Satz, obere Grenze ist die maximale Anzahl der Sätze im ausgewählten logischen Objekt. Bei der Angabe des Parameters ist zu bedenken, ob es sinnvoll ist, den Fokus durch eine große Anzahl von Sätzen zu bilden, weil Geschäftsbriefe im Durchschnitt kurz gefaßt sind und damit wenige Sätze enthalten.

Ein weiterer Parameter ist die Auswahl des Divisors mit dem die Normierung des Satzgewichtes durchgeführt wird. Wie schon oben erwähnt, kann zwischen zwei Möglichkeiten gewählt werden, entweder Division durch die Anzahl aller Wörter im Satz oder zweitens durch die Anzahl der Deskriptoren im Satz.

Beim Ermitteln des Fokus durch das INFOCLAS-Systems sollten folgende Probleme beachtet werden. Eine exakte Definition des Fokus ist nicht immer möglich. Einerseits muß bei der manuellen Bestimmung des Fokus durch mehrere Briefleser keine Übereinstimmung herrschen. Andererseits ist die Existenz eines Satzes mit der Kernaussage in einem Brief nicht zwingend bzw. die wichtigen Informationen des Briefes können über mehrere Sätze verteilt sein. Dies ist stark vom Schreibstil des

Briefautors abhängig und macht eine manuelle Fokussierung durch den Menschen schwierig und damit erst recht für den Computer.

Das zweite auf den Indexierer aufbauende Modul ist der Klassifizierer des INFOCLAS-Systems. Dieses Modul, das schon zur Vorklassifizierung bei der Fokussierung verwendet wurde, soll nun im nächsten Kapitel detailliert vorgestellt werden.

3.5. Klassifizierer

Drittes und letztes Modul des INFOCLAS-Systems ist das *Klassifizierungsmodul.* Bild 3.10 zeigt die Stellung des Klassifizierers im INFOCLAS-System.

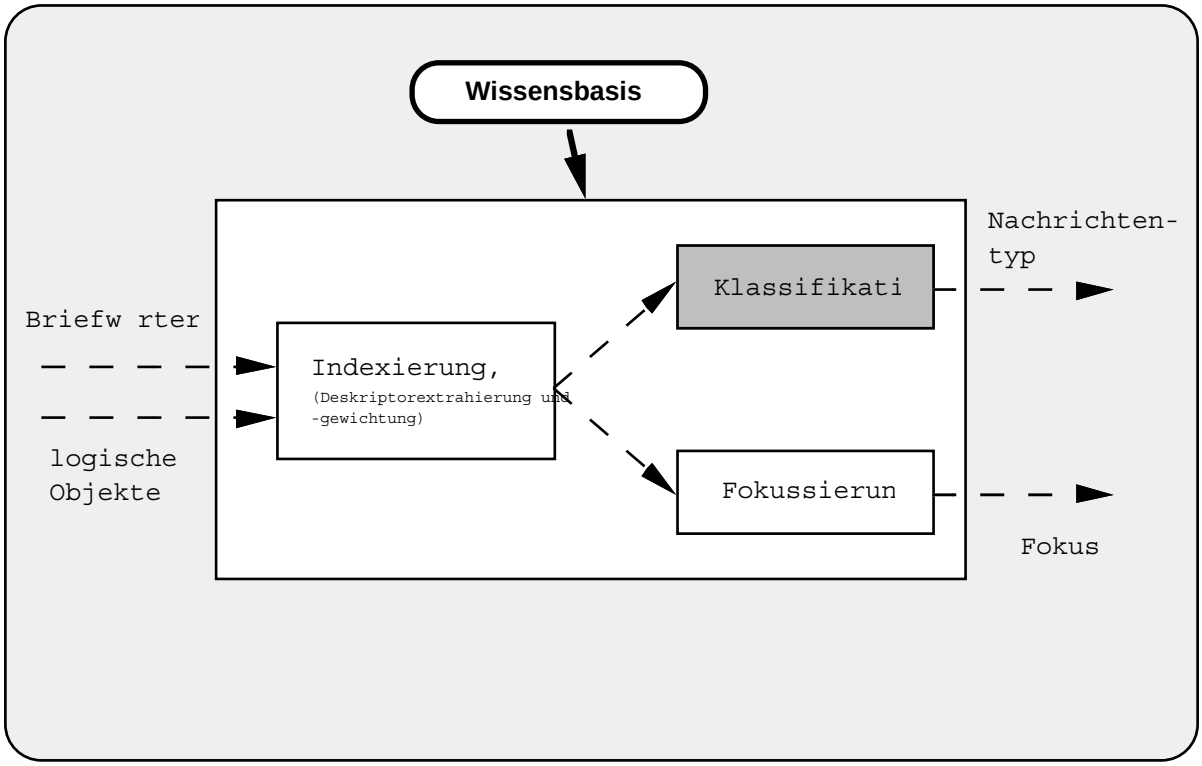


Bild **3.10**: INFOCLAS-System

Die Aufgabe des Klassifizierers ist die Aufstellung von Hypothesen über den Nachrichtentyp, in den der zu analysierende Brief einzuordnen ist. Momentan werden fünf verschiedene Nachrichtentypen vom System unterstützt (näheres siehe Kapitel 2.3.):

- Anfrage (ANFR);
- Angebot (ANGE);
- Bestellbestätigung (BEBE);
- Bestellung (BEST);
- Werbung (WERB).

Sollten im Verlaufe der Projektentwicklung weitere Nachrichtentypen modelliert werden, besteht eine einfache Möglichkeit das System dahingehend auszubauen. Die dafür durchzuführenden Maßnahmen werden in den Implementierungshinweisen (Abschnitt 4.1.2.) genauer beschrieben.

Repräsentiert werden die Nachrichtentypen durch *Primär-, Sekundär- und Tertiärwortlisten*. Unter Primär-, Sekundär- und Tertiärwortlisten sind die Mengen von Wörtern zu verstehen, die für einen bestimmten Nachrichtentyp spezifisch sind. Nun stellt sich die Frage, wie diese Wortmengen ermittelt wurden. Der Idealfall ist die Analyse von einer sehr großen Anzahl von Geschäftsbriefen eines bestimmten Nachrichtentyps. Bei dieser Analyse wird die Häufigkeit jedes einzelnen Wortstammes bestimmt und die Ergebnisse nach der Größe der Frequenz sortiert. Die an der Spitze dieser Liste liegenden Wortstämme bilden die Primärwortliste, die darauffolgenden Wortstämme die Sekundärwortliste, dann die Tertiärwortliste usw. Dies war bei der zur Verfügung stehenden Menge von Geschäftsbriefen (90) nicht so einfach möglich. Grundlage für die Bestimmung der einzelnen Primär-, Sekundär- und Tertiärwortlisten der einzelnen Nachrichtentypen waren berechnete Häufigkeitslisten, die aber manuell untersucht und aufgebessert werden mußten. Für eine Optimierung der Aussagekraft der Wortlisten, wäre der Einsatz von Synonymlexika und Thesauri sinnvoll. Hierdurch könnten die spezifischen Wortmengen der Nachrichtentypen umfangreicher und genauer spezifiziert werden.

Die Eingabe des Klassifizierers ist mit den vorhergehenden Modulen identisch, d.h. ein Brief wird von der Texterkennung analysiert und in das Doppellistenformat transformiert. Als Basis der Berechnung dient der Brief gemäß der INFOCLAS-internen Repräsentation (siehe Kapitel 3.3.) mit den ausgewählten Deskriptoren. Wiederum ist in diesem Modul eine Anzahl von Parametern für eine Berechnungsoptimierung einzustellen.

Der Klassifizierungsvorgang beginnt mit dem Einlesen eines Briefes. Dieser kann entweder neu von Datei oder aus der Wissensbasis geladen werden. Während des Ladevorgangs wird er in die interne Repräsentation umgewandelt. Vor der Berechnung der Briefklasse ist eine Bestimmung der Parameter notwendig. Die Parameter sind:

i) das logische Objekt (Basis der Berechnung);

ii) die Gewichtsfunktion (Deskriptorgewichtung);

iii) die Multiplikatoren (Modifikation der Deskriptorgewichte).

Mit dem ersten Parameter, dem logischen Objekt, wird der Bereich des Briefes definiert, aus dem die Deskriptoren zur Klassifizierung herangezogen werden. Um den Parameter festzulegen, existieren mehrere Möglichkeiten. Erstens kann jedes vorkommende logische Objekt angegeben werden, z.B. Briefrumpf, Betreffteil, Postscript usw. Zweitens können auch alle Deskriptoren des Briefes als Grundlage

gewählt werden durch Angabe, daß der ganze Brief das logische Objekt sei, d.h. der gesamte Brief wird intern als ein logisches Objekt aufgefaßt. Die letzte Möglichkeit der Parameterbelegung verbindet die ersten zwei Optionen. Hierbei wird mit jedem vorhandenen logischen Objekt eine Klassifizierung durchgeführt. Weiterhin erfolgt eine Klassifizierung mit Hilfe der Gesamtheit der Briefdeskriptoren. Die Ergebnisse der Berechnungen werden danach zu einem Gesamtresultat verrechnet. Diese Option könnte später um eine Gewichtung der einzelnen Objektergebnisse erweitert werden. Damit wäre eine Möglichkeit gegeben, die Resultate kleinerer logischer Objekte höher zu bewerten, z.B. den Betreffteil gegenüber dem Briefrumpf, weil in gewissen Fällen konkrete Angaben über die Klasse im Betreffteil vorkommen können.

Der zweite Parameter dient, wie beim Modul Fokussierer, zur Steuerung der Deskriptorgewichtung, durch Bestimmung einer geeigneten Gewichtsfunktion. Neben den bekannten Gewichtungskonzepten (Inverse Dokumenthäufigkeit, Informationswert, lokale Frequenz) kann auch jedem Deskriptor das Gewicht 1 zugeordnet werden, um eine Gleichgewichtung der Deskriptoren zu erreichen. Weitere Gewichtsfunktionen könnten aus dem Konzept der Diskriminanzwertbestimmung und aus Mischformen bestehender Funktionen gewonnen werden.

Mit dem dritten Parameter, den Multiplikatoren, wird die eigentliche Berechnung der Briefklasse gesteuert. Hierfür werden die nachrichtentypspezifischen Wortlisten verwendet. Das darauf aufbauende Klassifikationsverfahren hat folgenden Verlauf.

Für jeden der fünf Nachrichtentypen wird ein eigenes Gewicht errechnet. Diese Gewichte resultieren aus der Untersuchung der Bedeutung der einzelnen Deskriptoren für den jeweiligen Nachrichtentyp. Befindet sich ein Deskriptor in der Primärwortliste eines Nachrichtentyps, also in der Wortliste mit den wichtigsten und spezifischsten Wörtern des Typs, wird sein Gewicht mit dem Multiplikator *mul1* verrechnet. Er erhält dadurch ein höheres Gewicht und vergrößert damit das Gesamtgewicht des Nachrichtentyps für diesen Brief. Befindet sich der Deskriptor nicht in der Primärwortliste, so wird zuerst in der Sekundärwortliste und dann in der Tertiärwortliste gesucht. Bei einem Treffer wird entweder mit dem Multiplikator *mul2* verrechnet (Sekundärwortliste) oder mit dem Multiplikator *mul3* (Tertiärwortliste). Ist der Deskriptor in keiner der bisher erwähnten Wortlisten vertreten, wird sein Gewicht mit dem Multiplikator *mul4* entweder vermehrt oder vermindert. Damit wird gleichzeitig zum Ausdruck gebracht, daß die Multiplikatoren nicht immer zur Vergrößerung der Deskriptorgewichte dienen müssen, sondern auch zur Reduzierung. Wird als Wert für einen Multiplikator der Wert 0 gewählt, werden die entsprechenden Deskriptorgewichte bei der Berechnung der Gewichte der einzelnen Nachrichtentypen nicht mehr berücksichtigt. Die Nachrichtentypgewichte werden ermittelt durch Aufsummierung der Deskriptorgewichte, die sich aus dem Abgleich mit den Wortlisten ergeben haben.

Nach Bestimmung aller Nachrichtentypgewichte werden diese Maßzahlen größenmäßig sortiert, in Prozentwerte umgerechnet und in tabellarischer Form ausgegeben.

Eines der Probleme dieser Klassifizierung von Geschäftsbriefen ist die kleine Anzahl von fünf Nachrichtentypen. Die geringe Anzahl modellierter Nachrichtentypen ist in der zur Verfügung stehenden Testmenge von Geschäftsbriefen begründet. Um eine umfassendere Klassifizierung durchführen zu können, müßte eine größere Anzahl von Geschäftsbriefen zur Verfügung stehen und untersucht werden. In diesem Fall muß darauf geachtet werden, daß ein breites Spektrum von Geschäftsbriefarten vorhanden ist, damit die Menge der Nachrichtentypen sinnvoll erweitert werden kann. Weiterhin wäre bei einer größeren Anzahl von Geschäftsbriefen von einem speziellen Nachrichtentyp die Bestimmung der Primär-, Sekundär-, und Tertiärwortlisten leichter automatisch realisierbar. Eine weitere Verbesserung der Klassifikationsergebnisse ist durch die Verwendung von Synonymlexika und Thesauri innerhalb der Wortlisten möglich. Durch solche Hilfsmittel würden die Wortlisten der Nachrichtentypen genauer und umfangreicher spezifiziert.

Die bisherigen Wortlisten der Nachrichtentypen wurden teils automatisch, teils manuell erstellt. Dafür standen für jeden Nachrichtentyp nur zwischen 10 und 25 Briefe zur Verfügung. Man kann nicht davon ausgehen, daß aufgrund dieser Voraussetzungen eine ausreichende Spezifikation der Wortlisten gegeben ist. Diese könnte aber durch Anwendung der oben genannten Hilfsmittel und einer größeren Anzahl von Geschäftsbriefen erreicht werden.

4. Benutzerhandbuch

Im folgenden Kapitel wird die Handhabung des INFOCLAS-Systems beschrieben. Zuerst werden die Systemumgebung und der Aufbau der Dateistruktur beschrieben. Darunter fallen auch die Beschreibung der Dateien, die zum Laden des System notwendig sind. Anschließend erfolgt eine Auflistung der Funktionen, mit denen das System gestartet wird. Dabei geht man von einer dualen Bedienung für das System aus, d.h. es ist einerseits steuerbar durch einen menschlichen Benutzer, andererseits von anderen Programmexperten in einem größeren System anstoßbar. Dieses Konzept resultiert in einer menügesteuerten Oberfläche und in einer Schnittstelle bestehend aus automatisch aufrufbaren Funktionen.

4.1. Implementierungshinweise

4.1.1. Systemumgebung

Das innerhalb dieser Diplomarbeit entwickelte INFOCLAS-System wurde auf einer SUN-SPARCstation unter dem Betriebssystem UNIX [Bourne85] implementiert. Verwendetes Windowsystem ist OpenWindows, Version2. Der gesamte Programmcode wurde in Common Lisp mit einer CLOS (Common Lisp Object System) Erweiterung geschrieben [Steele91], [Keene89].

4.1.2. Dateistruktur und Laden des System

Der Programmcode des INFOCLAS-Systems ist über verschiedene Directories und Dateien verteilt. Die Dateistruktur und der Inhalt der Dateien wird nun im folgenden beschrieben. Zur Zeit der Systemimplementierung befanden sich alle Dateien in dem Directory *"/home/dittrich"*, auf SUN-SPARCstation. In der INFOCLAS-Dateistruktur existieren drei Arten von Dateien:

1) Dateien mit Programmcode;
2) Dateien für die Wissensbasis;
3) Dateien zum initialen Laden des Systems.

zu 1)

Inhalt der Dateien ist der Programmcode des INFOCLAS-Systems. Zur besseren Differenzierung der Systemteile wurde eine Unterteilung in vier Directories vorgenommen.

Directory: *infoclas*

Datei:	Inhalt
auto-funktionen	Funktionen zur programmgesteuerten Briefanalyse;
oberflaeche	INFOCLAS-Hauptmenü;

Directory: *indexierer*

Datei:	Inhalt
oberflaeche	Funktionen zur menügesteuerten Indexierung;
gewichtsfunktionen	Funktionen zur Berechnung der Deskriptorgewichte;
init	Initialisierungsfunktionen der Wissensbasis;
io	Funktionen zum Einlesen und Speichern von Briefdaten auf Dateien;
klasse	Funktion zur Definition der Klasse *letter*;
similarity	Funktionen zur Berechnung von Ähnlichkeitsmaßen;
hilfsfunktionen	Hilfsfunktionen für das INFOCLAS-System;

Directory: *fokussierer*

Datei:	Inhalt
oberflaeche	Funktionen zur menügesteuerten Fokussierung;
funktionen	Funktionen zur Berechnung der Fokushypothesen;

Directory: *klassifizierer*

Datei:	Inhalt
oberflaeche	Funktionen zur menügesteuerten Klassifizierung;
funktionen	Funktionen zur Berechnung der Nachrichtentyp-hypothesen;

zu 2)

Zur Anwendung der Systemfunktionalität benötigt INFOCLAS Daten aus der Wissensbasis. Diese sind in Dateien gespeichert, welche unter dem Directory *knowledgebase* zusammengefaßt sind. Die nachrichtentypspezifischen Wortlisten sind im Directory *nt-pool* zusammengefaßt.

Directory:	*knowledgebase*

Datei:	Inhalt
letter-number	enthält die aktuelle Briefnummer;
letters	enthält die bisher gespeicherten Briefe;
words	enthält die Deskriptorliste der Dokumentation (die in letters gespeicherten Briefe);

Directory:	*nt-pool*

Datei:	Inhalt
nt	enthält eine Liste mit den Namen der modellierten Nachrichtentypen;
words-ANFR	enthält die Primär-, Sekundär- und Tertiärwortliste des Nachrichtentyps **Anfrage**;
words-ANGE	enthält die Primär-, Sekundär- und Tertiärwortliste des Nachrichtentyps **Angebot**;
words-BEST	enthält die Primär-, Sekundär- und Tertiärwortliste des Nachrichtentyps **Bestellung**;
words-BEBE	enthält die Primär-, Sekundär- und Tertiärwortliste des Nachrichtentyps **Bestellbestätigung**;
words-WERB	enthält die Primär-, Sekundär- und Tertiärwortliste des Nachrichtentyps **Werbung**;

Folgende Maßnahmen im Directory *nt-pool* sind notwendig, um eine Erweiterung des Systems durch neu modellierte Nachrichtentypen durchzuführen:

a) Erweiterung der Liste in *nt* um die Abkürzungen der neuen Nachrichtentypen;

b Einrichtung einer Datei mit dem Namen *words-(neuer-Name)*, in der die Primär-, Sekundär- und Tertiärwortlisten des neuen Nachrichtentyps abgelegt sind.

zu 3)

Zum Laden des gesamten INFOCLAS-Systems muß die Datei *load-infoclas* in LISP aufgerufen werden.

Datei:	Funktion
load-infoclas	Laden des INFOCLAS-Systems;

Wegen der morphologischen Komponente MORPHIX wird dafür das Image *xpcl* benötigt. Um nur einzelne Systemmodule zu laden, existieren weiterhin folgende Dateien:

Datei:	Funktion
load-index	Laden des Indexierers;
load-focus	Laden des Fokussierers;
load-class	Laden des Klassifizierers;

Die automatische, programmgesteuerte Version des Systems kann nur durch das Laden des gesamten Systems bereitgestellt werden.

4.1.3. Starten des Systems

Da das INFOCLAS-System in zwei verschiedenen Modi Berechnungen durchführen kann, existieren zwei Möglichkeiten des Systemstartes. Um die erste Version einer menügesteuerten Oberfläche zu aktivieren, muß nach dem Ladevorgang die Funktion *start-infoclas* aufgerufen werden.

Funktionsname:	*start-infoclas*
Parameter:	keine;
Funktion:	Starten des Infoclas-Systems

(menügesteuerte Oberfläche);

Es existieren weiterhin Funktionen zum Starten einzelner Systemkomponenten. Diese Funktionen können aber nur dann aufgerufen werden, falls vorher die entsprechende Ladedatei aufgerufen wurde. Folgende Funktionen wurden dafür implementiert:

Funktionsname:	*start-index*
Parameter:	keine;
Funktion:	Starten des Indexierers (menügesteuerte Oberfläche);

Funktionsname:	*start-focus*
Parameter:	keine;
Funktion:	Starten des Fokussierers (menügesteuerte Oberfläche);

Funktionsname:	*start-class*
Parameter:	keine;
Funktion:	Starten des Klassifizierers (menügesteuerte Oberfläche);

Zur Anwendung der programmgesteuerten Analyse wurden die sogenannten *Auto-Funktionen* implementiert. Mit der Auto-Funktion *auto-infoclas* kann in diesem Modus das System gestartet werden. Die Beschreibung der einzelnen Funktionen, deren Parameter und Funktionsweise sowie ihre Aufrufreihenfolge erfolgt in Kapitel 4.3 *Programmgesteuerte Analyse.*

4.2. Menügesteuerte Analyse (Menü-Oberfläche)

Die menügesteuerte Oberfläche ist für menschliche Benutzer konzipiert, um die Indexier-, Fokussier- und Klassifiziervorgänge manuell zu steuern. Die verschiedenen Möglichkeiten der Parametrisierung erlauben dabei eine individuelle Anpassung des Systems an verschiedenartige Problemstellungen. Weiterhin ist ein leichtes Austesten der Verfahren durch Modifikation der Parameter gegeben. Die Resultate der Testläufe können anschließend für die Grundeinstellungen der automatischen Schnittstelle des Systems benutzt werden.

Nach dem Laden und Starten des INFOCLAS-Systems erscheint auf dem Monitor das "INFOCLAS - Hauptmenü".

INFOCLAS	
Indexierer	(1)
Fokussierer	(2)
Klassifizierer	(3)
INFOCLAS beenden	(q)

Bild **4.1**: INFOCLAS-Hauptmenü

Durch Auswahl einer der drei Module beginnt ein interner Ladevorgang, bei dem Daten aus der Wissensbasis bereitgestellt werden. Durch Aufbau eines Auswahlmenüs kann die Bearbeitung mit Hilfe des ausgewählten Moduls beginnen. Erwähnt sei in diesem Zusammenhang, daß die Menüs, welche eine Option mit dem Buchstaben "b" (back) besitzten, damit in Richtung des vorhergehenden Menüs verlassen werden können. Die Menüs mit einem Buchstaben "q" (quit) beenden das ausgewählte Modul bzw. das gesamte System. Bei Verlassen eines Moduls wird jeweils die gesamte Wissensbasis auf die entsprechenden Dateien gespeichert.

Bei der Eingabe von Werten für die Parameter werden vom System interne Überprüfungen der Korrektheit durchgeführt. Diese Korrektheitstests werden aber in den folgenden Abschnitten nicht immer beschrieben.

4.2.1. Die Oberfläche des Indexierers

Hat man das Modul Indexierer ausgewählt erscheint das "Auswahl-Indexierer"-Menü.

Auswahl-Indexierer	
- Brief bearbeiten	(1)
- Wissensbasis	(2)
- Indexierer beenden	(q)

Bild **4.2**: Auswahl-Indexierer

Mit Hilfe des Menüs, welches in gleicher Form in jedem Modul existiert, erhält der Benutzer die Möglichkeit in die Wissensbasis zu verzweigen oder eine Briefbearbeitung auszuführen.

a) Wissensbasis

In der Wissensbasis stehen unter anderem die bisher gespeicherten Briefe und die Deskriptoren aus diesen Briefen zur Verfügung. Der einzelne Brief ist mit allen acht, früher berechneten Instanzenslots abgespeichert. Die Deskriptoren der Dokumentation (Menge der gespeicherten Briefe) liegen in der Deskriptorliste vor. Jeder Deskriptor wird durch die Stammform des Ausgangswortes, die Häufigkeit des Auftretens dieser Stammform in der Dokumentation und einer Liste der Briefe in dem der Deskriptor vorkommt, repräsentiert.

Hat man sich im Auswahlmenü für die Wissensbasis entschieden, kann entweder die Wissensbasis neu initialisiert oder Informationen über in der Wissensbasis befindliche Daten ausgeben werden.

Wissensbasis	
- Initialisieren	(1)
- Ausgabe	(2)
- Zurück zu: Auswahl-Indexierer	(q)

Bild **4.3**: Wissensbasis

Durch den Befehl "Initialisieren" werden alle Briefe und die Deskriptorliste gelöscht und die Grundeinstellungen vorgenommen.

Sollen Informationen ausgegeben werden, besteht eine Auswahl unter folgenden Daten:

i) aktuelle Briefnummer, d.h. die Nummer des nächsten neu eingelesenen Briefes;

ii) Anzahl der Briefe in der Wissensbasis;

iii) die Namen aller in der Wissensbasis gespeicherten Briefe, mit den zugehörigen Briefnummern;

iv) die Anzahl der Deskriptoren in der Deskriptorliste;

v) die Liste aller Briefe (Sloteinträge);

vi) die Deskriptorliste mit zugehörigen Informationen.

b) Briefbearbeitung

Mit der Briefbearbeitung betritt der Benutzer den Teil der Module, durch den Briefe neu eingelesen und analysiert werden können. Um zwischen schon in der Wissenbasis gesicherten und neu einzulesenden Briefen wählen zu können, erfolgt zuerst die Abfrage danach durch nachstehendes Menü:

Auswahl der Briefart	
- Neuen Brief einlesen	(1)
- Brief aus der Wissensbasis	(2)
- Zurück zu: Auswahl-Indexierer	(b)

Bild **4.4**: Auswahl der Briefart

Wird ein neuer Brief eingelesen, muß ein Briefname und der Name der Datei angegeben werden, in dem sich der Brief im Doppellistenformat befindet. Der Name der Datei muß dabei der Pfadname im Stringformat sein. Weiterhin muß entschieden werden, ob die Wissensakquisitionskomponente der morphologischen Analyse ein- oder ausgeschaltet sein soll. Bei eingeschalteter Komponente werden Wörter, die noch nicht in den Lexika von MORPHIX gespeichert sind, auf linguistische Eigenschaften abgefragt. Diese Eigenschaften werden dann in Verbindung mit dem Wort in MORPHIX abgelegt.

Ist der zu analysierende Brief schon in der Wissensbasis vorhanden, so genügt nach Auswahl des entsprechenden Menüpunktes eine Eingabe des Briefnamens.

Nachdem der Einlesevorgang beendet ist, wählt man die Bearbeitungsart des Briefes aus:

Auswahl der Bearbeitungsart	
- Ausgabe Brief	(1)
- Deskriptorgewichtung	(2)
- Brief sichern	(3)
- Brief löschen	(4)
- Zurück zu: Auswahl der Briefart	(b)

Bild **4.5**: Auswahl der Bearbeitungsart

Mit den Funktionen, die an die einzelnen Menüpunkte gebunden sind, können folgende Bearbeitungsarten durchgeführt werden:

i) Ausgabe von Briefdaten;

ii) Gewichtung der Deskriptoren;

iii) Sichern und Löschen von Briefen.

zu i)

Mit Hilfe der Ausgabeoption ist der Benutzer in der Lage, Informationen über den Brief abzufragen. Das Ausgabemenü der Briefslots enthält 9 Kategorien.

Ausgabe Briefslots	
- Alle Slots	(1)
- Name	(2)
- Nummer	(3)
- Logische Objekte	(4)
- Sätze	(5)
- Wörter	(6)
- Wortliste (ges.)	(7)
- Wortliste (morph.)	(8)
- Wortliste (stop. & morph.)	(9)
- Zurück zu: Ausgabe Brief	(b)

Bild **4.6**: Ausgabe Briefslots

Folgende Informationen sind hiermit abrufbar:

1) alle Sloteinträge des aktuellen Briefes;
2) der Briefname;
   Bsp.: "NAME:"
   "BRIEF48"
3) die Briefnummer;
   Bsp.: "NUMMER:"
   27
4) Liste der logischen Objekte mit der Gesamtanzahl;
   Bsp.: "LOG. OBJEKTE:"
   (3 (SUBJECT BODY ENCLOSURE))
5) Liste mit der Gesamtanzahl der Sätze und der Anzahl in den jeweiligen logischen Objekten;
   Bsp.: "SÄTZE:"
   (5 ((SUBJECT 1) (BODY 3) (ENCLOSURE 1)))
6) Liste mit der Gesamtanzahl der Wörter und der Anzahl in den jeweiligen logischen Objekten;
   Bsp.: "WÖRTER:"
   (81 ((SUBJECT 14) (BODY 63) (ENCLOSURE 4)))

7) Repräsentation des Briefes als Liste mit Unterlisten für jedes logische Objekt und wiederum Unterlisten für jeden Satz mit vorgestellter Satznummer; Wörter und Satzzeichen sind als Strings aufgeführt;

Bsp.: "WORTLISTE-ALL:"
(((1
("Betr." ":" "Vertrag-Nr." "5-112-4308" "ueber" "\"" "Anfertigung" "einer" "Studie" """ "Wissensbank" """ """ """)))
((1
("Sehr" "geehrte" "Damen" "und" "Herren" "," "als" "Anlage" "uebersenden" "wir" "Ihnen" "den" "Entwurf" "des" "obigen" "Vertrages" "in" "dreifacher" "Ausfertigung" "mit" "der" "Bitte" "um" "Unterzeichnung" "und" "Ruecksendung" "von" "zwei" "Exemplaren" "an" "die" "DLR" "," "Hauptabteilung" "Beschaffung" "," "Linder" "Hoehe" "," "5000" "Koeln" "90"".))
(2
("Nach" "Gegenzeichnung" "durch" "uns" "erhalten" "Sie" "ein" "Original" "des" "Vertrages" "fuer" "Ihre" "Akten" "."))
(3
("Mit" "freundlichen" "Gruessen" "i.A." "R" "Derkum")))
((1
("Anlagen" ":" "3" "Vertragsexemplare"))))

8) Listenstruktur wie in 7), aber ohne Satzzeichen sowie einer morphologischen Analyse derWörter, d.h. die Wörter sind durch ihre Stammform ersetzt;

Bsp.: "WORTLISTE-MORPH:"
(((1
("betr" "vertrag-nr" "5-112-4308" "ueber" "\"" "anfertigung" "ein" "studie" "wissensbank")))
((1
("sehr" "ehr" "dame" "und" "herr" "als" "anlage" "uebersend" "wir" "sie" "d-" "entwurf" "d-" "obig" "vertrag" "in" "dreifach" "ausfertigung" "mit" "d-" "bitt" "um" "unterzeichnung" "und" "ruecksendung" "von" "zwei" "exemplar" "an" "d-" "dlr" "hauptabteilung" "beschaffung" "linder" "hoehe" "5000" "koeln" "90"))
(2
("nach" "gegenzeichnung" "durch" "wir" "erhalt" "sie" "ein" "original" "d-" "vertrag" "fuer" "ihr" "akt"))
(3
("mit" "freundlich" "gruss" "i" "r" "derkum")))
((1
("anlage" "3" "vertragsexemplar"))))

9) Liste der Briefdeskriptoren; für jeden Deskriptor eine Liste mit Stammform, Wortart und Frequenz innerhalb des Briefes;

Bsp.: "WORTLISTE-RED:"
(("akt" MORPHIX:NOMEN 1)
("anfertigung" MORPHIX:NOMEN 1)
("anlage" MORPHIX:NOMEN 2)
("ausfertigung" MORPHIX:NOMEN 1)
("beschaffung" MORPHIX:NOMEN 1)
("bitt" MORPHIX:VERB 1)
("dame" MORPHIX:NOMEN 1)
("dreifach" MORPHIX:ADJEKTIV 1)
("ehr" MORPHIX:ATTRIBUTIV-GEBRAUCHTES 1)
("entwurf" MORPHIX:NOMEN 1)
("erhalt" MORPHIX:VERB 1)
("exemplar" MORPHIX:NOMEN 1)
("freundlich" MORPHIX:ADJEKTIV 1)
("gegenzeichnung" MORPHIX:NOMEN 1)
("gruss" MORPHIX:NOMEN 1)
("hauptabteilung" MORPHIX:NOMEN 1)
("herr" MORPHIX:NOMEN 1)
("hoehe" MORPHIX:NOMEN 1)
("obig" MORPHIX:ADJEKTIV 1)
("original" MORPHIX:NOMEN 1)
("ruecksendung" MORPHIX:NOMEN 1)
("studie" MORPHIX:NOMEN 1)
("uebersend" MORPHIX:VERB 1)
("unterzeichnung" MORPHIX:NOMEN 1)
("vertrag" MORPHIX:NOMEN 2)
("vertragsexemplar" MORPHIX:NOMEN 1)
("wissensbank" MORPHIX:NOMEN 1))

Eine weitere Ausgabemöglichkeit ist die Suche nach den Positionen eines bestimmten Wortes im Brief. Nach der Abfrage des Wortes wird in Listenformat ausgegeben, in welchen Sätzen und Wortposition das Wort im Brief auftritt.

zu ii)

Mit der zweiten Bearbeitungsart, der Deskriptorgewichtung, wird das Hauptverfahren des Indexierers aktiviert. Zur Auswahl der Gewichtungsfunktion erscheint nachstehendes Menü:

Welche Gewichtsfunktion erwünscht ?	
- Frequenz	(1)
- Inverse Dokumenthäufigkeit	(2)
- Informationswert	(3)
- Diskriminanzwert	(4)
- Erweiterungen	(5)
- Mischformen	(6)
- Zurück zu: Bearbeitungsart	(b)

Bild **4.7**: Auswahl der Gewichtsfunktion

Durch die Wahl einer dieser Gewichtsfunktionen ordnet das System den ausgewählten Deskriptoren Gewichtskennzahlen zu. Bei allen angebotenen Gewichtsfunktionen muß vor Ausführung der Gewichtsberechnung geklärt werden, ob ein einzelnes Wort oder alle Wörter des Briefes gewichtet werden sollen. Bei Gewichtung eines einzelnen Wortes wird dies vorher abgefragt.

Bei der Gewichtsfunktion "Frequenz" kann unterschieden werden, ob die Berechnung sich auf den aktuellen Brief oder die Dokumentation bezieht.

Bei vollständiger Integrierung der Gewichtsfunktion *Diskriminanzwert* muß vor Bestimmung der Gewichte mit Hilfe eines Auswahlmenüs eine Ähnlichkeitsfunktion gewählt werden, mit der die Berechnungen durchgeführt werden. Folgende Ähnlichkeitsfunktionen stehen zur Auswahl:

a) Cosinuskoeffizient;

b) Dicekoeffizient;

c) Jaccardkoeffizient.

Zu einigen der Gewichtsfunktionen ist eine Erweiterung der Berechnung möglich. Diese kann durch Auswahl des Menüpunktes "Erweiterung" eingestellt werden. Die bisher realisierte Erweiterung basiert auf einem Abgleich der Deskriptoren mit den Primär- und Sekundärwortlisten der modellierten Nachrichtentypen. Bei einem Treffer wird das Deskriptorgewicht mit dem zugeordneten Multiplikator verrechnet. Dafür können mit Hilfe des "Erweiterungsmenüs" Voreinstellungen vom Benutzer eingegeben werden.

Erweiterungsmenü	
- Einschalten	(1)
- Multiplikator für Primärwörter	(2)
- Multiplikator für Sekundärwörter	(3)
- Ausgabe Multiplikatoren	(4)
- Ausschalten	(5)
- Zurück zu Erweiterungen	(b)

Bild **4.8**: Erweiterungsmenü

Durch "Einschalten" und "Ausschalten" wird die Zuschaltung der Erweiterung gesteuert. Die Zahlenwerte der einzelnen Multiplikatoren sind individuell veränderbar, müssen aber positiv sein. Zur Überprüfung der aktuellen Erweiterungsparameter ist eine Visualisierung mittels "Parameterausgabe" möglich.

Ausgabe der Deskriptorgewichtung ist eine Tabelle mit den Wörtern in der ersten Spalte und den Gewichten in der zweiten Spalte. Im Beispiel (Bild 4.9) wird die Ausgabe einer Indexierung des Briefs Nr. 48 gezeigt, mit der Gewichtsfunktion *Inverse Dokumenthäufigkeit* unter Berücksichtigung aller Deskriptoren des Briefes.

```
*******************************************************************************
Inverse Dokumenthäufigkeit der Worte aus dem Brief BRIEF48:
-------------------------------------------------------------------------------
                akt                     5.39
                anfertigung             5.39
                anlage                  3.87
                ausfertigung            5.39
                beschaffung             5.39
                bitt                    1.69
                dame                    4.39
                dreifach                5.39
                ehr                     1.07
                entwurf                 5.39
                erhalt                  2.39
                exemplar                3.39
                freundlich              1.07
                gegenzeichnung          5.39
                gruss                   1.07
                hauptabteilung          5.39
                herr                    1.14
                hoehe                   5.39
                obig                    5.39
                original                5.39
                ruecksendung            4.39
                studie                  4.39
                uebersend               4.39
                unterzeichnung          5.39
                vertrag                10.78
                vertragsexemplar        5.39
                wissensbank             5.39
*******************************************************************************
```

Bild **4.9**: Ausgabe der Fokussierungshypothesen des Briefes Nr. 48

zu iii)

Durch Sichern und Löschen eines Briefes wird eine Verwaltung der Briefe in der Datenbasis realisiert. Ein Brief, der aus der Wissensbasis geladen wurde, kann dabei nicht noch einmal gesichert werden.

4.2.2. Die Oberfläche des Fokussierers

Zur Aktivierung des Fokussierungsmoduls muß im INFOCLAS-Hauptmenü der Punkt 2 ausgewählt werden (Bild 4.1). Wie im Modul Indexierer besteht auch hier eine Verzweigung zur Wissensbasis oder zur eigentlichen Briefbearbeitung.

a) Wissensbasis

Die Optionen und Menüs sind identisch mit denen im Indexierungsmodul (siehe 4.2.1.a)).

b) Briefbearbeitung

Das Einlesen eines Briefes von Datei oder aus der Wissensbasis erfolgt nach dem gleichen Schema, welches schon im Indexierer angewendet wurde (vgl. 4.2.1 b)). Nach dem Einlesen des Briefes erscheint wiederum ein Briefbearbeitungsmenü:

Auswahl der Bearbeitungsart	
- Ausgabe Brief	(1)
- Fokussierung	(2)
- Indexierung	(3)
- Brief sichern	(4)
- Brief löschen	(5)
- Zurück zu: Auswahl der Briefart	(b)

Bild **4.10**: Auswahl der Bearbeitungsart

Die Einträge, *Briefausgabe*, *Brief sichern* und *Brief löschen* sind identisch mit Einträgen im Indexierungsmodul (vgl. auch 4.2.1, Bild 4.5). Mit dem Eintrag *Indexierung* wird das Gewichtsfunktionenmenü des Indexierers und somit seine Funktionalität aufgerufen.

Mit *Fokussierung* wird der eigentliche Fokussierungsvorgang aktiviert und führt zu folgendem Menü:

Fokussierungsmenü	
- Log. Objekt (def.: body)	(1)
- Gewichtsfunktion (def.: inv. Dok.)	(2)
- Anzahl der Sätze (def.: 1)	(3)
- Divisor bei Gewichtung (def.: all)	(4)
- Parameterausgabe	(5)
- Start Fokussierung	(6)
- Zurück zu: Briefbearbeitung	(b)

Bild **4.11**: Fokussierungsmenü

An dieser Stelle der Briefbearbeitung ist der Benutzer in der Lage, die Parameter für die Fokussierung auszugeben bzw. neu zu definieren oder die Berechnung des Fokus zu starten. Zur Fokussierung sind vier Parameter nötig:

1) das logische Objekt, welches die Basis für die Fokusberechnung bildet (Deskriptormenge) (Voreinstellung: Briefrumpf);
2) die Gewichtungsfunktion, mit der den Deskriptoren des ausgewählten logischen Objekts ein Gewicht zugeordnet wird (Voreinstellung: Inverse Dokumenthäufigkeit);
3) die Anzahl der Sätze, die der Fokus umfassen soll (Voreinstellung: 1);
4) der Divisor, mit dem die Satzgewichte normiert werden (Voreinstellung: Anzahl aller Wörter im Satz).

zu 1)

Zur Festlegung eines logischen Objekts stehen in einer Auswahl alle logischen Objekte zur Verfügung, die dem INFOCLAS-System von diesem Brief zugänglich gemacht wurden (Doppellistenformat). Ein einzelnes logisches Objekt kann direkt durch Angabe des Namens bestimmt werden. Eine weitere Option ist die Wahl des gesamten Briefs als logisches Objekt (*ALL*), d.h. alle Deskriptoren des Briefes werden für die Berechnung herangezogen. In diesem Zusammenhang sei erwähnt, daß wenn vom gesamten Brief die Rede ist, nur die logischen Objekte und damit deren Wörter gemeint sind, die dem System zur Verfügung stehen. Im Regelfall ist dies nicht der komplette Geschäftsbrief wie er auf Papier realisiert ist mit Empfängeradresse, Firmenlogo usw. (weniger für statistische Verfahren geeignet).

Mit Hilfe des Menüeintrages *merge* (m) können beliebige logische Objekte zu einem Objekt zusammengefaßt werden. Die ausgewählten logischen Objekte werden intern als ein Objekt angesehen und behandelt. Bei der Auswahl ist es nicht möglich, ein logisches Objekt doppelt anzugeben. Die Defaulteinstellung für das logische Objekt

ist "body", da man davon ausgehen kann, daß jeder Brief einen Rumpf besitzt und hier die meisten Informationen über den Briefinhalt stehen.

Welches logische Objekt ?	
- SUBJECT	
- BODY	
- ALL	
- MERGE	(m)
- Zurück zu: Fokussierungsmenü	(b)

Bild **4.12**: Auswahl des logischen Objekts (Beispielbrief)

zu 2)

Für die Auswahl der Gewichtsfunktion stehen die gleichen Funktionen zur Verfügung, die schon im Indexierer vorgestellt wurden.

Welche Gewichtsfunktion erwünscht ?	
- Frequenz	(1)
- Inverse Dokumenthäufigkeit	(2)
- Informationswert	(3)
- Diskriminanzwert	(4)
- Erweiterungen	(5)
- Mischformen	(6)
- Zurück zu: Fokussierungsmenü	(b)

Bild **4.13**: Auswahl Gewichtungsfunktion

Soll eine Erweiterung der Gewichtsfunktion zugeschaltet werden, so ist gegenüber der Indexierung eine weitere Möglichkeit vorhanden. Neben dem Abgleich der Wörter mit den Wortlisten der Nachrichtentypen (Deskriptor in NT-Pool) ist eine *Vorklassifizierung* ausführbar.

Erweiterung zur Gewichtsfunktion	
- Vorklassifizierung	(1)
- Deskriptor in NT-Pool	(2)
- Zurück zu: Gewichtsfunktionen	(b)

Bild **4.14**: Erweiterungen zur Gewichtsfunktion

Diese Vorklassifizierung führt eine interne Klassifikation des Briefes durch. Dafür werden Defaulteinstellungen der Parameter verwendet, wobei der Briefrumpf als logisches Objekt genommen wird. Das Resultat der Vorklassifizierung ist genau der Nachrichtentyp, der das größte Vertrauensmaß erhalten hat. Die restlichen Hypothesen werden in der Vorklassifizierung nicht berücksichtigt.

zu 3)

Auch die Fokusgröße ist individuell einstellbar. Dafür wird dem Benutzer das aktuelle logische Objekt mit der Anzahl der Sätze ausgegeben. Die Eingabe muß dann eine ganze Zahl zwischen 1 und der maximalen Satzanzahl sein.

Die Defaulteinstellung für die Fokusgröße ist ein Satz. Wird nach der Festlegung der Fokusgröße das logische Objekt gewechselt, findet eine Überprüfung der Verträglichkeit statt und es erfolgt eine Fehlermeldung, falls der Fokus für das neue logische Objekt zu groß sein sollte.

zu 4)

Als Divisor für die Normierung der Satzgewichtung stehen zwei Alternativen zur Verfügung:

a) die Anzahl aller Wörter eines Satzes;

b) die Anzahl aller Deskriptoren eines Satzes.

Die Defaulteinstellung ist die Alternative a).

Alle obigen Parameter sind zur Überprüfung des aktuellen Zustands mit der Option *Parameterausgabe* visualisierbar (Menüeintrag (5), Bild 4.11).

Die Aktivierung der Fokusbestimmung wird mit *Start Fokussierung* durchgeführt, die dann mit den aktuellen Parametern abläuft. Werden keine Parameter neu eingestellt, so läuft die Bestimmung mit den Defaulteinstellungen ab. Bei Verlassen des Menüs *Fokussierung*, z.B. um einen neuen Brief einzulesen, werden die Parameter wieder auf die Ausgangswerte eingestellt.

Die Ausgabe der Fokussierung erfolgt in einer Tabelle mit den angeführten Werten pro Spalte: Nummer der Wahrscheinlichkeit, Satznummer oder Nummern der Sätze in der Satzgruppe, Wert der Hypothese (Satz- oder Satzgruppengewicht). Im Beispiel (Bild 4.15) wurde der Brief48 verwendet und folgende Parametereinstellungen vorgenommen:

logisches Objekt: body;

Gewichtsfunktion: Inverse Dokumenthäufigkeit + Vorklassifizierung;

Fokusgröße: ein Satz

```
**********************************************************
Fokushypothesen in absteigender Reihenfolge
**********************************************************
                1. Satz: 2      2.44
                2. Satz: 1      2.28
                3. Satz: 3      0.36
**********************************************************
```

Bild **4.15**: Ausgabe der Fokussierungshypothesen des Briefes Nr. 48

4.1.3. Die Oberfläche des Klassifizierers

Der dritte Menüpunkt des INFOCLAS-Hauptmenüs startet das Modul Klassifizierer. Die Struktur der Oberfläche ist an die der zwei anderen Module angeglichen und in gewissen Bereichen identisch. Wiederum findet eine Zweiteilung zwischen Wissensbasis und Briefbearbeitung statt.

a) Wissensbasis

Gleiche Struktur und Aufbau wie im Indexierer (siehe 4.1.1. a)).

b) Briefbearbeitung

Der Einlesevorgang, die Briefausgabe, das Löschen und Sichern eines Briefes sowie die Indexierung sind identisch mit den gleichnamigen Funktionen im Modul Fokussierer (siehe 4.1.2.). Wählt man im Menü "Auswahl der Bearbeitungsart" des Klassifizierers den zweiten Menüpunkt erscheint das Klassifikationsmenü.

Klassifikationsmenü	
- Logisches Objekt (def.: body)	(1)
- Gewichtsfunktion (def.: inv. Dok.)	(2)
- Multiplikator	(3)
- Parameterausgabe	(4)
- Start Klassifikation	(5)
- Zurück zu: Bearbeitungsart	(b)

Bild **4.16**: Klassifizierungsmenü

Wie im Fokussierer sind auch hier mehrere Parameter für die Klassifizierung einstellbar. Diese Parameter sind:

1) das logische Objekt, welches die Basis für die Klassenberechnung bildet (Deskriptormenge) (Voreinstellung: Briefrumpf);

2) die Gewichtungsfunktion, mit der den Deskriptoren des ausgewählten logischen Objekts ein Gewicht zugeordnet wird (Voreinstellung: Inverse Dokument-häufigkeit);

3) die Multiplikatoren, mit denen die Gewichte der Deskriptoren verrechnet werden (Voreinstellung: mul1 = 1.5, mul2 = 1.3, mul3 = 0.0, mul4 = 0.0);

zu 1)

Zur Auswahl des logischen Objekts und damit der berücksichtigten Deskriptoren steht jedes logische Objekt einzeln sowie der ganze Brief (all) zur Verfügung. Eine weitere Möglichkeit ist der Menüpunkt "all and single". Damit sind mehrere Klassifizierungen verbunden sowohl auf dem einzelnen logischen Objekten als auch auf dem gesamten Brief. Die einzelnen Ergebnisse werden dann zu einem Gesamtergebnis verrechnet.

Die Defaulteinstellung ist das logische Objekt "body".

Welches logische Objekt ?	
- SUBJECT	
- BODY	
- ALL	(a)
- ALL and Single	(s)
- Zurück zu: Fokussierungsmenü	(b)

Bild **4.17**: Auswahl des logischen Objekts
(Beispielbrief)

zu 2)

Für den Parameter der Gewichtsfunktion stehen die im Modul Indexierer vorgestellten Funktionen zur Verfügung, erweitert um die Möglichkeit, keine Gewichtsfunktion anzugeben. Hierbei wird jedem Deskriptor das Gewicht 1 zugeordnet.

Defaulteinstellung ist die Gewichtungsfunktion "inverse Dokumenthäufigkeit".

Welche Gewichtsfunktion erwünscht ?	
- Keine	(0)
- Frequenz	(1)
- Inverse Dokumenthäufigkeit	(2)
- Informationswert	(3)
- Diskriminanzwert	(4)
- Mischformen	(5)
- Zurück zu: Klassifikationsmenü	(b)

Bild **4.18**: Auswahl der Gewichtsfunktion

zu 3)

Insgesamt sind vier Multiplikatoren zur Berechnung der Klasse vom Benutzer festzulegen. Der erste Multiplikator wird mit den Gewichten der Deskriptoren verrechnet, die in den Primärwortlisten der modellierten Nachrichtentypen vorkommen. Der zweite Multiplikator mit denen in den Sekundärwortlisten, der dritte Multiplikator mit denen in den Tertiärwortlisten. Mit dem vierten Multiplikator werden die Deskriptoren verrechnet, die in keiner der Wortlisten vorkommen.

Die über Tests ermittelten Defaulteinstellungen der einzelnen Multiplikatoren sind:

Multiplikator1 (mul1): 1.5;

Multiplikator2 (mul2): 1.3;

Multiplikator3 (mul3): 0.0;

Multiplikator4 (mul4): 0.0.

Die Klassifikation wird durch Auswahl von *Start Klassifikation* (Bild 4.16) aktiviert und mit den eingestellten Parametern durchgeführt. Die Ausgabe des Resultats erfolgt in einer Tabelle mit den angeführten Werten pro Spalte: Abkürzung des Nachrichtentyps, Prozentangabe und interner Wert der Klassifikation in Klammern. Diese Ausgabe ist sortiert nach der Wahrscheinlichkeit der Hypothese. Im Bild 4.19 wird die Ausgabe einer Klassifikation des Briefes NR.48 gezeigt, die mit den

Voreinstellungen der Parameter durchgeführt wurde. Die internen Werte in den Klammern stellen die Nachrichtentypgewichte dar und sind die direkte Grundlage für die Prozentangaben.

```
*********************************************
      Klassifizierungshypothesen für
          den Brief BRIEF48:
---------------------------------------------------
BEBE:        41.81 %        (71.92)
ANGE:        19.52 %        (33.59)
ANFR:        15.14 %        (26.05)
BEST:        14.86 %        (25.57)
WERB:         8.66 %        (14.90)
*********************************************
```

Bild **4.19**: Ausgabe der Klassifizierungshypothesen des Briefes Nr. 48

4.3. Programmgesteuerte Analyse (Auto-Funktionen)

Neben der durch eine Menüoberfläche von einem Benutzer steuerbaren Version des INFOCLAS-Systems wird es durch spezielle Funktionsaufrufe ermöglicht, die Funktionalität des Systems ohne Oberfläche zu nutzen. Bei diesen Funktionen handelt es sich um Schnittstellenfunktionen, bei denen die Parameter, die bei der Oberflächenversion des Systems mit Menüs einzeln eingelesen wurden, im Funktionsaufruf unmittelbar angegeben werden. Eine Ausnahme besteht darin, wenn die Defaulteinstellung eines Parameters benutzt werden soll.

Im folgenden Abschnitt werden deshalb die sogenannten "Auto"- Funktionen mit ihren Parametern vorgestellt. Gibt es für einen Parameter eine Auswahl von Möglichkeiten, so wird diese bei der Erläuterung des Parameters angegeben. In diesem Zusammenhang sei erwähnt, daß für diese Art der Systemnutzung lediglich rudimentäre Fehlerabfangmechanismen implementiert wurden und deshalb auf korrekte Angabe der Parameterwerte geachtet werden muß. Weiterhin gilt für alle Auto-Funktionen, daß die Briefinstanz, die von *auto-read-letter* erzeugt wird, in der globalen Variable *letter* abgelegt wird. Somit ist der Wert, der für *Briefinstanz* angegeben werden muß, die globale Variable *letter*.

Um die Funktionalität des Systems zu nutzen, müssen zuerst Daten aus der Wissensbasis geladen werden. Damit wird der Grundzustand des Systems hergestellt. Dieses wird durch die Funktion "auto-infoclas" gewährleistet.

Funktionsname:	*auto-infoclas*
Parameter:	keine;
Funktion:	Herstellung des Systemgrundzustands;

Das System muß mit der Funktion "auto-stop-infoclas" beendet werden, damit eine korrekte Sicherung der Wissensbasis erfolgen kann.

Funktionsname:	*auto-stop-infoclas*
Parameter:	keine;
Funktion:	Beendigung des Systems und Sicherung der Wissensbasis in Dateien;

Gemäß Kapitel 4.1. wird die Zweiteilung in Wissensbasis und Briefbearbeitung beibehalten.

4.3.1. Wissensbasis

Zur Verwaltung der Wissensbasis werden Funktionen zur Verfügung gestellt, mit denen die Wissensbasis initialisiert, ein Brief gesichert oder gelöscht werden kann.

Funktionsname:	*auto-init-kb*
Parameter:	keine;
Funktion:	initialisiert die Wissensbasis;

Funktionsname:	*auto-save-letter*
Parameter:	- *inst* Briefinstanz;
Funktion:	sichert den Brief in der Wissensbasis;

Funktionsname:	*auto-delete-letter*
Parameter:	- *name* Briefname;
Funktion:	löscht den Brief aus der Wissensbasis;

Informationen über die Inhalte der Wissensbasis können über die globalen Variablen *letter-number*, *letters-kb* und *words-kb* abgefragt werden, die nach Starten des Systems mit der Funktion "auto-infoclas" mit Daten belegt sind.

4.3.2. Briefbearbeitung

Um eines der drei Module des INFOCLAS-Systems anwenden zu können, muß zuerst ein Brief geladen und in die interne Datenstruktur umgewandelt werden. Dafür muß die Funktion *auto-read-letter* aufgerufen werden. Mit dem Parameter *letter-new* wird gesteuert, ob der Brief aus einer Datei oder der Wissensbasis geladen wird. Durch diese Funktion wird der globalen Variablen *letter* eine Instanz der Klasse *letter* zugewiesen, deren Slots mit den Werten des eingelesenen Briefes gefüllt werden.

Funktionsname:	*auto-read-letter*	
Parameter:	- *name*	Name des Briefes;
	- *pathname*	Pfadname der Briefdatei;
	- *letter-new*	(default: nil) Argumente: **NIL**: Brief aus der Wissensbasis **T**: Brief aus der Datei mit dem Pfadnamen *pathname*;
	- *aqui*	(default: nil) Ein- bzw. Ausschalten der Morphix-Wissensakquisitionskomponente
Funktion:	liest Brief ein und erzeugt Briefinstanz;	

Ein Zugriff auf die einzelnen Sloteinträge der Briefinstanz erfolgt durch die nachstehenden objekt-orientierten Zugriffsfunktionen:

Funktionsname	Funktionsart	Slot
letter-name	reader	Briefname
letter-num	reader	Briefnummer
get-log-object	accessor	logische Objekte
get-sents	accessor	Sätze
get-words	accessor	Wörter
get-wordlist-all	accessor	Briefgesamtliste
get-wordlist-morph	accessor	morph. Briefgesamtliste
get-wordlist-red	accessor	Deskriptorliste

Die Ausführung eines speziellen INFOCLAS-Moduls wird jeweils durch eine spezielle Auto-Funktion geleistet. Die verschiedenen Variationsmöglichkeiten innerhalb der Module wird über Funktionsparameter gesteuert.

1) Indexierer

Funktionsname:	*auto-index*	
Parameter:	- *inst*	Briefinstanz; in *letter* abgelegt
	- *nt-ext* (default: nil)	Erweiterungen ein/aus;
	- *mulp*	(default: 1) Primärmultiplikator;
	- *muls*	(default: 1) Sekundärmultiplikator;
	- *weight-fun*	(default: invers-doc-weight) Gewichtsfunktion; Argumente: - invers-doc-weight - information-weight - frequenz
Funktion:	gewichtet die Briefdeskriptoren;	

Die Ausgabe der Auto-Indexierung ist eine Liste mit Gewichten, die den Deskriptoren paarweise zugeordnet werden müssen. Im folgenden Beispiel wird die Ausgabe eines Auto-Indexiervorgangs gezeigt. Verwendet wurden dafür die Defaultparameter und der Beispielbrief Nr. 48.

Beispielausgabe des Auto-Indexierers:

```
(("akt" 5.392318S0)
("anfertigung" 5.392318S0)
("anlage" 3.865772S0)
("ausfertigung" 5.392318S0)
("beschaffung" 5.392318S0)
("bitt" 1.691878S0)
("dame" 4.392318S0)
("dreifach" 5.392318S0)
("ehr" 1.07039S0)
("entwurf" 5.392318S0)
("erhalt" 2.392318S0)
("exemplar" 3.392318S0)
```

("freundlich" 1.07039S0)
("gegenzeichnung" 5.392318S0)
("gruss" 1.07039S0)
("hauptabteilung" 5.392318S0)
("herr" 1.14439S0)
("hoehe" 5.392318S0)
("obig" 5.392318S0)
("original" 5.392318S0)
("ruecksendung" 4.392318S0)
("studie" 4.392318S0)
("uebersend" 4.392318S0)
("unterzeichnung" 5.392318S0)
("vertrag" 10.78464S0)
("vertragsexemplar" 5.392318S0)
("wissensbank" 5.392318S0))

2) Fokussierer

Funktionsname:	auto-focus	
Parameter:	- *inst*	Briefinstanz; in *letter* abgelegt
	- *log-object*	(default: (body)) Liste der gewünschten logischen Objekte;
	- f-weight-fun	(default: inv) Gewichtsfunktion; Argumente: -inv - inf - freq
	- *sent*	(default: 1) Fokusgröße in Sätzen (sentences);
	- *divisor*	(default: all-words) Satzgewichtnormierung; Argumente: - all-words - keywords
	- *pre-class*	(default: nil) Vorklassifizierung ein/aus;
	- *nt-ext*	(default: nil) Erweiterungen ein/aus;

	- *mulp*	(default: 1) Primärmultiplikator;
	- *muls*	(default: 1) Sekundärmultiplikator;
	- *mulc1*	(default: 1.5) Primärmultiplikator für Vorklassifizierung;
	- *mulc2*	(default: 1.3) Sekundärmultiplikator für Vorklassifizierung;
Funktion:	liefert Fokushypothesen	

Die Ausgabe der Auto-Fokussierung ist eine Liste der Satznummern denen Satzgewichte zugeordnet sind. Verwendet wurden dafür die Defaultparameter und der Beispielbrief Nr. 48.

Beispielausgabe des Auto-Fokussierers:

((2 2.257993S0)
(1 2.059602S0)
(3 0.3567966S0))

3) Klassifizierer

Funktionsname:	auto-class	
Parameter:	- *inst*	Briefinstanz; in *letter* abgelegt
	- *log-object*	(default: body) logisches Objekt;
	- *weight-fun*	(default: invers-doc-weight) Gewichtsfunktion; Argumente: - invers-doc-weight - information-weight - frequenz - no-weight
	- *mul1*	(default: 1.5) Primärmultiplikator;

- *mul2* (default: 1.3)
  Sekundärmultiplikator;
- *mul3* (default: 0.0)
  Tertiärmultiplikator;
- *mul4* (default: 0.0)
  Restwortmultiplikator;

Funktion: liefert Klassifikationshypothesen

Die Ausgabe der Auto-Klassifizierung ist eine sortierte Liste mit Nachrichtentyp-hypothesen denen Prozentangaben und die internen Werte zugeordnet sind. Verwendet wurden dafür die Defaultparameter und der Beispielbrief Nr. 48.

Beispielausgabe des Auto-Klassifizierers:

```
((BEBE    39.742509651184086    32.341278261265401)
(ANGE     27.86555633544922     22.676165125442189)
(BEST     23.393360352516176    19.036824379414551)
(ANFR     21.671823906898503    17.635888965086178)
(WERB     10.211532878875733    8.3098432687916848))
```

5. Ergebnisse und Tests

Im folgenden Abschnitt werden Erfahrungen mit dem INFOCLAS-System und Testergebnisse vorgestellt. Dabei wird einzeln auf die drei Module des Systems eingegangen.

Für das gesamte System gilt, daß durch eine Vergrößerung der Datenbasis eine starke Verlangsamung der Vorgänge *Speichern* und *Löschen* eines Briefes eintritt. Dies ist durch die Such- und Erweiterungsmaßnahmen innerhalb der Deskriptorliste begründet.

5.1. Indexierer

Im Modul Indexierer wurden besonders die Gewichtsfunktionen getestet, die nun einzeln bewertet werden.

a) Frequenz

   Diese Gewichtsfunktion ermöglicht eine sehr schnelle Berechnung der Deskriptorgewichte, da alle Berechnungen schon beim Ladevorgang des Briefes durchgeführt werden. Die Aussagekraft der Gewichte ist aber sehr beschränkt, weil die Berechnungen sich nur auf den aktuellen Brief beziehen und keinen Bezug zur Dokumentation herstellen.

b) Inverse Dokumenthäufigkeit

   Auch für diese Gewichtsfunktion werden entscheidende Vorberechnungen beim Laden des Briefes ausgeführt, so daß eine schnelle Ermittelung der Deskriptorgewichte ermöglicht wird. Die Berechnungsdauer steigt zwar mit dem Anwachsen der Dokumentation, aber nicht so stark wie bei später diskutierten Gewichtsfunktionen. Die Aussagekraft der Ergebnisse ist größer als bei der Funktion *Frequenz*, da Informationen aus der Dokumentation bei der Berechnung berücksichtigt werden. Nach den durchgeführten Tests lieferte diese Gewichtsfunktion das beste Verhältnis zwischen Zeitaufwand und Aussagekraft der Ergebnisse. Deshalb wurde diese Funktion bei der menügesteuerten Analyse und den Auto-Funktionen als Voreinstellung gewählt.

c) Informationswert

   Bei dieser Gewichtsfunktion wird eine noch stärkere Einbindung der Dokumentation bei der Gewichtsberechnung vollzogen (Stichwort: Ballast). Dieses bedingt eine längere Berechnungsdauer und auch eine stärkere Abhängigkeit von der Größe der Dokumentation. Die Ergebnisse sind als ähnlich gut zu bezeichnen wie die der *Inversen Dokumenthäufigkeit*.

d) Diskriminanzwert

Diese Gewichtsfunktion wurde nicht zur Gänze in das System eingebunden, sondern nur rudimentär getestet. Vor einer vollständigen Einbindung wurde bisher abgesehen, da mit ihr ein sehr hoher Berechnungsaufwand verbunden wäre, welcher sehr stark mit der Größe der Dokumentation ansteigen würde. Sollten die Datenstrukturen der Datenbasis und die Suchvorgänge verbessert werden, sind mit guten Ergebnissen dieser Gewichtsfunktion zu rechnen. Es ist aber zu diesem Zeitpunkt nicht abzusehen, ob die Ergebnisse einen solchen Rechen- und Zeitaufwand rechtfertigen [Willett85].

5.2. Fokussierer

Die in den durchgeführten Tests gelieferten Resultate sind schwer zu bewerten, da es selbst für einen menschlichen Leser schwer ist, den Fokus eines Briefes zu bestimmen. Einige Feststellungen konnten aber doch getroffen werden. Erstens tritt eine Verbesserung der Ergebnisse ein, wenn eine Vorklassifizierung des Briefes ausgeführt wird. Es ist aber zu bedenken, daß die Gefahr besteht, bei einer falschen Klassifizierung eine noch stärkere Verfälschung des Ergebnisses zu bekommen. Zweitens wäre eine Reduzierung der Fokusgröße auf Phrasen oder Deskriptorbereiche sinnvoll, auch mit Hinsicht auf die Schwierigkeiten der Ermittlung der Satzzeichen während der Phase der Texterkennung.

5.3. Klassifizierer

Bei der Klassifizierung konnten erfolgreichere Tests durchgeführt werden. Es hat sich dabei herausgestellt, daß eine Klassifizierung aufgrund der Deskriptoren aus dem logischen Objekt *body* oder dem gesamten Brief am sinnvollsten ist. Für die Werte der Multiplikatoren der nachrichtentypspezifischen Wortlisten wurden folgende Richtlinien ermittelt. Mit den Multiplikatoren für die Primär- und Sekundärwörter soll eine Erhöhung der Deskriptorgewichte erfolgen. Die Gewichte der Tertiär- und Restwörter werden am besten vernachlässigt oder zumindest stark vermindert. Die Gewichtsfunktion mit den besten Ergebnissen ist die *Inverse Dokumenthäufigkeit* (Begründung siehe Abschnitt 5.1).

Abschließend werden zwei Testläufe vorgestellt, die die Klassifizierungsgüte mit Hilfe der Gewichtungsfunktionen *Inverse Dokumenthäufigkeit* und *Informationswert* aufzeigen sollen.

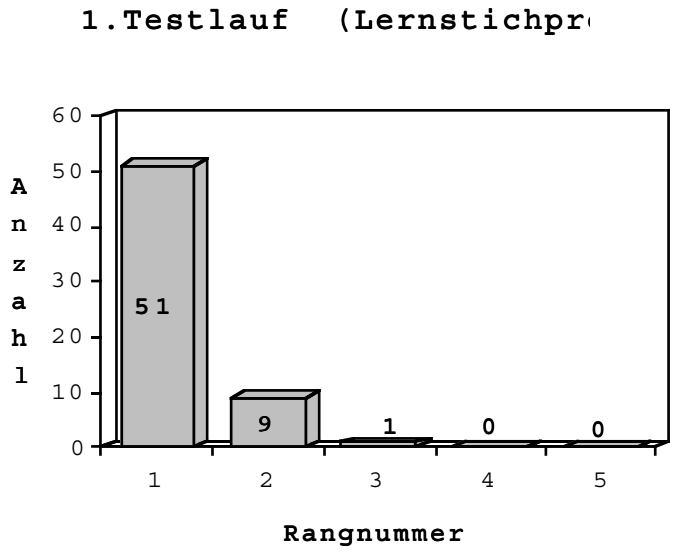
An den X-Achsen der Graphiken sind die jeweiligen Rangnummern der korrekten Nachrichtentypen bei der Erkennung abgetragen, an den Y-Achsen die Anzahl der Briefe mit dem jeweiligen Ergebnis.

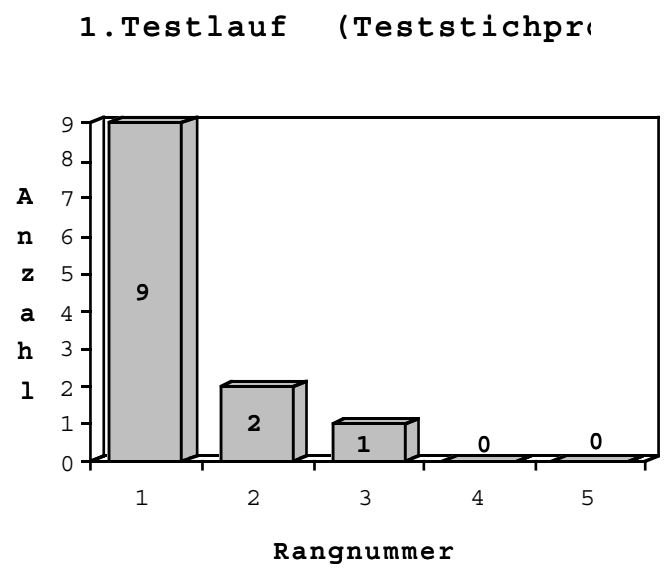
1. Testlauf

Der Testlauf wurde in zwei Phasen durchgeführt. In der ersten Phase wurden die Briefe klassifiziert, die in der Lernstichprobe enthalten sind und zu den fünf modellierten Nachrichtentypen gehören. In der zweiten Phase wurden 12, dem System unbekannte, neue Briefe untersucht (Teststichprobe).

Die bei der Klassifikation verwendeten Parameter hatten folgende Werte:

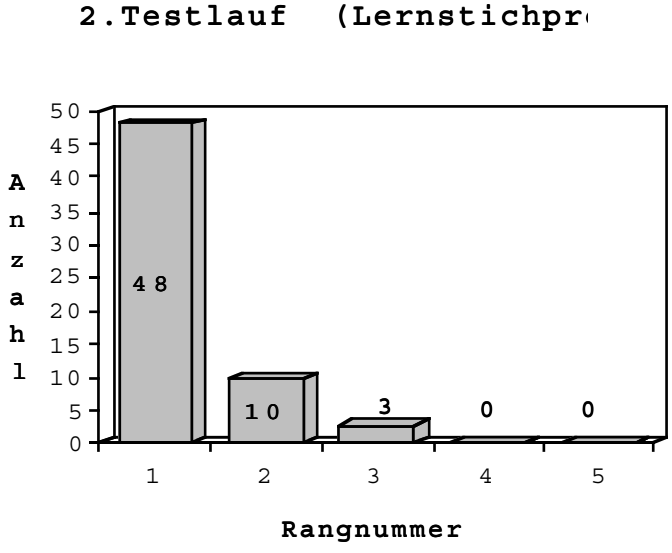
logisches Objekt:	body;
Gewichtsfunktion:	Inverse Dokumenthäufigkeit;
mul1:	1.5;
mul2:	1.3;
mul3:	0.0;
mul4:	0.0;

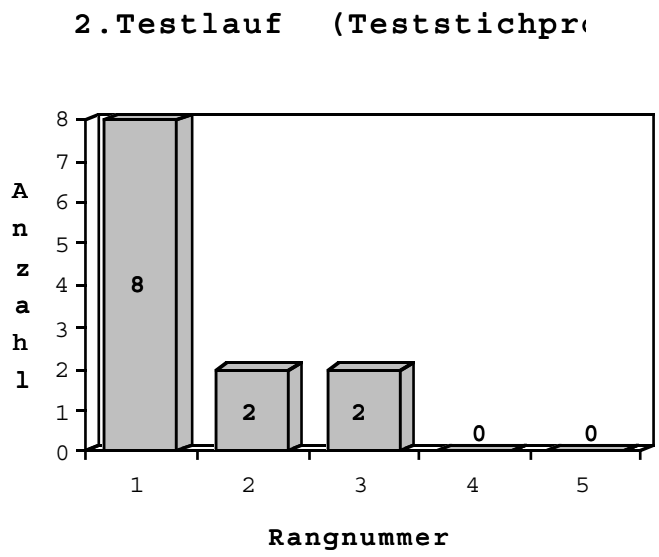




2.Testlauf

Auch im zweiten Testlauf wurde in zwei Phasen getestet (Lernstichprobe /Teststichprobe). Die verwendeten Parameter sind identisch mit Ausnahme der Gewichtsfunktion. Zur Gewichtung wurde der Informationswert herangezogen.





Aus den durchgeführten Tests ist ersichtlich, daß im Rahmen der vorliegenden Grundlagen (20 Briefe umfassende Dokumentation, 5 modellierte Nachrichtentypen) gute Klassifikationsergebnisse erzielt wurden. Dabei beträgt der Prozentanteil der richtig erkannten Nachrichtentypen für die Gewichtsfunktion *Inverse Dokumenthäufigkeit* ≈ 84% bei der Lernstichprobe und 75% bei der Teststichprobe. Etwas schwächer sind die Ergebnisse bei der Gewichtsfunktion *Informationswert*; hier ist der Wert für die Lernstichprobe ≈ 79% und für die Teststichprobe ≈ 66%. Zu beachten ist, daß für ein aussagekräftigeres Ergebnis bei den Teststichproben eine größere Anzahl von neuen Geschäftsbriefen nötig wäre.

6. Verwandte Ansätze aus der Dokumentanalyse (WISDOM Projekte)

6.1. Einleitung

Innerhalb dieses Abschnitts werden Klassifikationswerkzeuge aus dem Verbundprojekt WISDOM (Wissensbasierte Systeme zur Bürokommunikation: Dokumentbearbeitung, Organisation, Mensch-Computer Kommunikation) [Lutze91] beschrieben. Dieses Verbundprojekt lief von 1986 bis 1989 unter Federführung der TA Triumph-Adler AG in Zusammenarbeit mit der Gesellschaft für Mathematik und Datenverarbeitung mbH (GMD), der Universität Stuttgart, der Frauenhofer-Gesellschaft und anderen Instituten.

Im folgenden werden die Konzepte einzelner Systeme aus diesem Verbundprojekt vorgestellt. Speziell wird dabei auf die Methoden und Verfahren zur Erkennung von wichtigen und relevanten Textteilen für die Klassifikation eingegangen. Diese Identifizierung von Textkomponenten soll zur Unterstützung der Entwicklung von Hypothesen über die Art und den Typ des zu bearbeitenden Dokuments dienen.

6.2. EPIKUR-System

Das erste System, das vorgestellt wird und innerhalb des *WISDOM*-Verbundprojekts entwickelt wurde, ist das *EPIKUR*-System [Kreplin86]. Dabei handelt es sich um ein Expertensystem zur Postbearbeitung in Kommunikations- und Retrievalsystemen. Mit diesem Analyse- und Klassifikationssystem soll sowohl elektronisch angekommene als auch papiergebundene Post automatisch analysiert und verteilt werden. Letztere wird vorher mit einem Scanner eingelesen und damit dem System zugänglich gemacht.

Die drei Hauptfunktionen des entwickelten Postexperten sind:

i) Verteilung,

ii) Sortierung,

iii) Archivierung

der Postdokumente. Diese Funktionen können durch individuelle Angaben des Benutzers gesteuert werden.

Unter *Verteilung* wird die Weitergabe von zentral eingegangener Post auf die einzelnen Empfänger verstanden, wobei man zwischen persönlich adressierter und nicht-persönlich adressierter Post unterscheidet wird. Bei letzterer Art muß automatisch entschieden werden, wer aufgrund seines Tätigkeitsbereichs oder Funktion eine Zustellung erhält.

Bei der *Sortierung* wird auf Basis von individuellen Angaben des Benutzers eine automatische Unterteilung der Post in Kategorien wie wichtige Post, weniger wichtige

Post, Werbung, dringende Post, weiterzuleitende Post an Mitarbeiter usw. vorgenommen.

Archivierung ist das automatische Ablegen der Post in ein Archiv aufgrund von inhaltlichen Verweisen aus den Dokumenten.

Aufbau und Arbeitsweise des EPIKUR-Systems werden hauptsächlich von einer Inferenzkomponente bestimmt. Mit ihr wird erstens die formale und inhaltliche Analyse des Dokuments durchgeführt, die ein Dokumentenprofil als Ergebnis hat. Zweitens wird eine Klassifizierung des Dokuments aufgrund vorgegebener Klassentypen durchgeführt. Dabei stützt man sich auf eine Wissensbasis, in der Weltwissen und Klassenbeschreibungen von Dokumentarten gespeichert sind. Bei der Formalanalyse wird das Layout des Dokuments untersucht und daraus Hinweise für die Entwicklung des Dokumentenprofils, welches dieses Dokument beschreibt, gezogen. Bei der inhaltlichen Analyse werden aus dem Dokumenttext inhaltstragende Komponenten herausgefiltert und weiterverarbeitet. Diese inhaltstragenden Komponenten sind:

i) Deskriptoren (gewichtete Einzelwörter);

ii) Konzepte (semantische Konstrukte aus Begriffen);

iii) Begriffe (Tokenmengen, welche die Wertemenge für die in der Wissensbasis gespeicherten Prädikate bilden).

Die *Deskriptoren* werden nach dem Verfahren des *Fuzzy-Retrievals* [Salton87] ermittelt. Hierbei werden Schlüsselwörter nach ihrer Bedeutung für das Dokument gewichtet. Damit können Dokumente mit *Ähnlichkeitsmaßen* untereinander verglichen werden. Die Gewichte werden nach den klassischen *Gewichtsfunktionen* [Salton87] berechnet.

Um *Konzepte* innerhalb des Dokuments erkennen zu können, werden diese mit Hilfe einer Beschreibungssprache dargestellt. Diese Beschreibungssprache besteht aus Prädikaten über Layout-, Logik-, Synonym- und Linguistikbeziehungen. Die Prädikate, mit denen die Konzepte beschrieben werden können, müssen vorher in die Wissensbasis eingetragen werden. Es können also nur die Konzepte erkannt werden, die auch in der Datenbasis bekannt sind.

Ein Beispiel für zwei Konzepte sind Angebot und Autor eines Artikels:

```
Angebot <=
        SEMANTISCHE_BEZIEHUNG (Synonym(Angebot) and
        (instance_of(Produkt)) or (instance_of (Dienstleistung)))
```

```
Artikel_Autor <=
        STEHT_AUF_1._SEITE (instance_of(Person))
                nicht_gefunden:
        STEHT_AUF_1._SEITE
                        (instance_of(Behörde) or instance_of(Firma))
```

Mit der dritten Art inhaltstragender Komponenten, den *Begriffen*, werden die Prädikate gefüllt und dadurch ausgewertet. Die Wörter, die die Deskriptoren, Konzepte und Begriffe bilden, werden vorher mit Hilfe von sprachabhängigen, morphologischen Deflexionsalgorithmen auf ihre Grundformen (Nominativ Singular bzw. Infinitiv) abgebildet.

Unter Anwendung von Zuordnungsverfahren wird das erzeugte Dokumentenprofil mit den Klassendefinitionen verglichen und ein Typ festgelegt. Diese Verfahren bedienen sich dabei gängigen Ähnlichkeitsmaßen, z.B. dem *Cosinusmaß*. Das Dokumentenprofil wurde vorher aus den Ergebnissen der formalen und inhaltlichen Analyse gebildet.

In der Wissensbasis, die zum Teil vom Benutzer gefüllt werden muß und zum Teil vorgegeben ist, befindet sich Wissen über:

- Thesaurusrelationen (z.B. Ober-/Unterbegriff, Synonym);
- (gesicherte) wissensmodellierende Relationen (z.B. instance_of);
- Wissen über den bestehenden und früheren Zustand;
- Layoutwissen (z.B. steht_auf_1.Seite);
- Wissen über semantische Zusammenhänge (z.B. part_of).

6.3. MULTOS-System

Ähnlich wie im EPIKUR-System verläuft die Erkennung von relevanten Komponenten in Dokumenttexten im *MULTOS*-System [Lutz89] zielgerichtet. Das heißt, es werden Prädikate für zu erkennende Komponenten generiert, die mit Begriffen gefüllt und verifiziert werden müssen.

Ziel des Klassifikationswerkzeugs innerhalb des MULTOS-Systems ist wiederum die Zuordnung des Dokuments zu einem Dokumenttyp. Diese Dokumente können elektronisch oder durch Scanner eingelesene papiergebundene Texte sein. Um aber im Gegensatz zum EPIKUR-System eine abstraktere Sicht zu bekommen, wird beim MULTOS-System eine Dokumenttyphierarchie in der Wissensbasis abgespeichert. Dabei wird ein spezielles Dokumentenmodell [Barbic&Rabitti85] angewandt. In diesem Modell wird für jedes Dokument ein Baum aufgebaut, der *konzeptuelle Strukturdefinition* (CSD) genannt wird. Die Knoten des Baumes werden als Konzeptkomponenten bezeichnet und kommen in drei Arten vor:

1.) *Basis-Komponenten*, welche grundsätzlich Blattknoten (Terminalknoten) repräsentieren und auf relevante Textteile im Dokument verweisen. Diese können Wörter, Sätze, Abschnitte etc. sein;

2.) *Komplex-Komponenten*, aus diesen Knoten kann durch Aggregation nach bestimmten Konstruktionsregeln wie Auswahl, Sequenz und Wiederholung eine weitere Verfeinerung der CSD durchgeführt werden;

3.) *Spring-Komponenten*, die für eine spätere Verfeinerung innerhalb der Struktur der Typhierarchie vorgesehen sind. Damit wird das Prinzip des schwachen Typs (weak type) realisiert.

Weitere Einzelheiten über das Dokumentenmodell und CSD finden sich in [Barbic85], [Eirund88], [Eirund&Kreplin88] und [Lutz89].

Beispiel für eine CSD des Typs "Angebot":

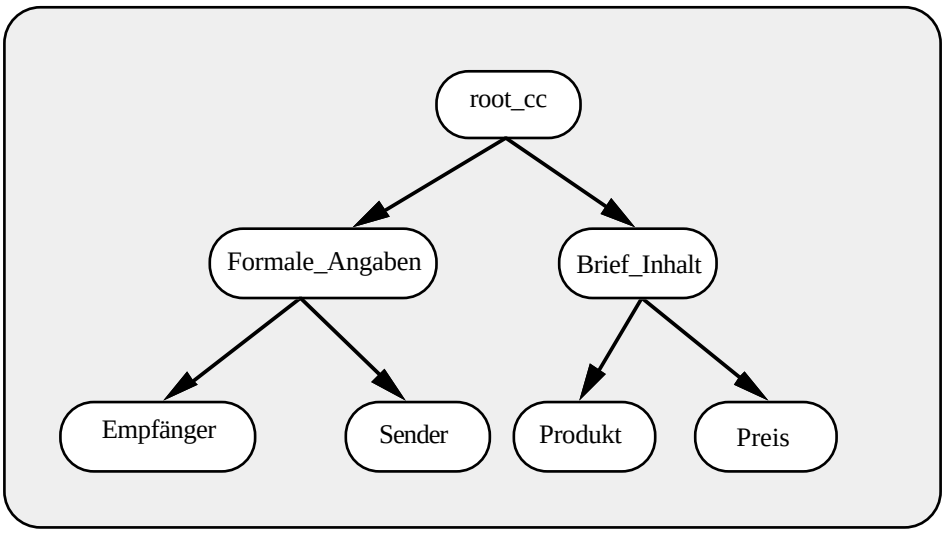


Bild **6.1**: CSD-Struktur, vom Typ "Angebot"

Der Dokumenttyp eines bearbeiteten Dokuments wird durch Festlegung des Typs mit der größten und feinsten Übereinstimmung aus der Typhierarchie bestimmt. Dies wird durch einen schrittweisen Durchlauf durch die Hierarchie vom Wurzelknoten bis zum entsprechenden Typ erreicht (Entscheidungsbaum). Es wird also eine schrittweise Verfeinerung durchgeführt.

Um nun die relevanten Dokumentteile für die einzelnen Komponenten der CSD zu finden, wird Wissen in Form von CDL-Sätzen (Content Description Language) [Eirund87] angegeben. CDL ist eine Beschreibungssprache ähnlich der, die im EPIKUR-

System zum Einsatz kommt. Die einzelnen Prädikate in der CDL sind von folgender Art:

- Prädikate, die Layout- und logische Elemente beschreiben (z.B. Tabellen, Zeilen);
- Positionsoperatoren für Layout- und Logikelemente (z.B. nach, erstes);
- Prädikate für Schlüsselwort- und Tokenmatch;
- Semantische Relationen zu Hintergrundwissen (z.B. instance_of);
- Grammatikalische Relationen (z.B. possessiv, temporal);
- Referenzen zu anderen konzeptuellen Komponenten;
- Positionsoperatoren für Schlüsselwörter, Relationen und Referenzen (z.B. nach, zwischen);
- Kombinationen von Prädikaten, die Mengen- und Logikoperatoren enthalten.

Für jede konzeptuelle Komponente der CSD wird ein CDL-Satz generiert und zugewiesen, wobei auch innere Knoten der CSD mit CDL-Sätzen verbunden sind. Dies wird dazu benutzt, den Suchraum für relevante Textteile von Subkomponenten einer Komponente einzuschränken. Zum Beispiel im Bereich des Dokumenttextes, der den Briefinhalt beinhaltet, liegen auch diejenigen Textkomponenten, welche für die Subkomponenten <Produkt> und <Preis> relevant sind. Damit wird eine Einschränkung des Suchraums für relevante Inhaltsteile von der Wurzel bis zu den Blättern der CSD erreicht.

Nachfolgend ein Beispiel für eine Angebot-CSD mit CDL-Sätzen.

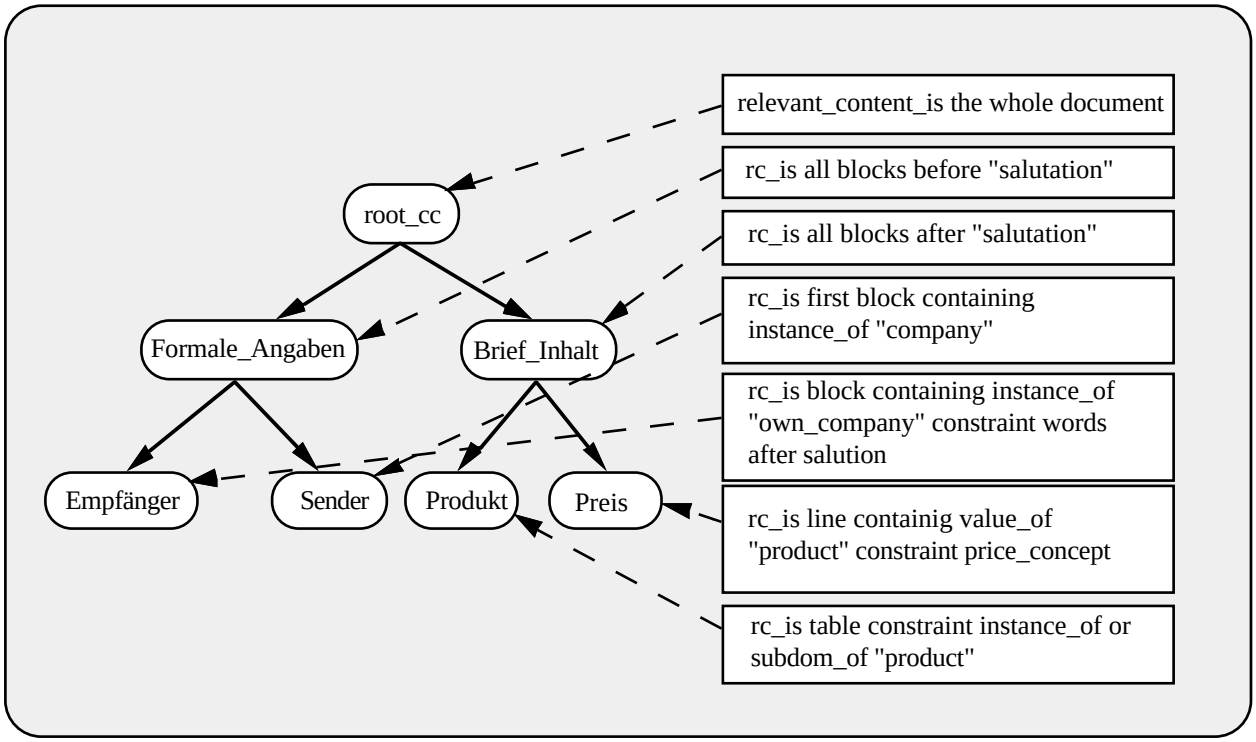


Bild **6.2**: CSD für Dokumenttyp "Angebot“ mit CDL-Sätzen

Mit der CDL wird festgelegt in welcher Form, Aussehen, Art, Position etc. die Dokumentteile innerhalb des Dokuments vorhanden sein müssen. Hierbei wird also der logische Aufbau und die Layoutstruktur des Dokuments berücksichtigt. Weiterhin werden Schlüsselwörter aus dem Text und spezielle Begriffe, wie Produktnamen, Namen, Preise usw. bearbeitet und in die CDL-Sätze eingesetzt. Die Schlüsselwörter und Begriffe werden vorher über einen speziellen Präprozessor einer Morphologie-analyse unterzogen.

Weiterhin steht eine Wissensbasis zur Verfügung. Das darin gespeicherte Wissen läßt sich in drei unterschiedliche Klassen einteilen:

i) Strukturelles Wissen

ii) Hintergrundwissen

iii) Aktuelles Wissen.

Unter *strukturellem Wissen* versteht man die Typhierarchie, die Typ-CSD´s jedes einzelnen Typs in der Typhierarchie sowie die CDL-Prädikate, die die konzeptuellen Komponenten spezifizieren und den Suchraum der dafür relevanten Textteile vorgeben.

Unter *Hintergrundwissen* werden Angaben über die Domäne des betrachteten Dokuments (Organisationsnamen, Produktnamen, Mitarbeiter usw.) sowie Synonyme und weiteres linguistisches Wissen verstanden.

Im Bereich des *aktuellen Wissens* stehen die Angaben über das aktuell bearbeitete Dokument, d.h. seine Repräsentation als Ganzes, die Layoutstruktur, die logische Struktur und die ermittelten Wörter mit ihren morphologischen Grundformen.

6.4. WAK-Projekt

In einem Teilprojekt des WISDOM-Verbundprojekts, dem WAK-Projekt [Rieder87], wurde ein Klassifikationswerkzeug entwickelt, das ebenfalls auf dem MULTOS-Dokumentenmodell [Barbic&Rabitti85] basiert. Die bisher entworfenen Klassifikatoren benutzen folgende Funktionen:

- Wortvergleich;
- Morphologiekomponente;
- Suchbereichseinschränkung für relevante Inhaltsteile;
- Tabellenanalysator;
- sowie eine Wissensbasis.

Um auch inhaltstragende Elemente aus natürlichsprachlichen Dokumenttexten zu extrahieren, wurde folgender Ansatz innerhalb des WAK-Projekts gewählt. Es sollen komplex strukturierte Konzepte mit oder ohne explizit im Text ausgedrückten Beziehungen zwischen den einzelnen Elementen identifiziert werden, z.B.:

> "... bieten Ihnen die Portierung des Programms X von Rechner A auf Rechner B an."

Bei diesem Beispiel sollte korrekterweise erkannt werden, daß nicht das Programm oder der Rechner sondern vielmehr eine Portierung angeboten wird.

Um die Arten der komplex strukturierten Konzepte und ihrer Vorkommen in Dokumenttexten aufzuzeigen, wurde eine manuelle Untersuchung von 48 Geschäftsbriefen vom Typ "Angebot" durchgeführt. Ein Teil der CSD-Struktur für den Typ "Angebot" sieht folgendermaßen aus:

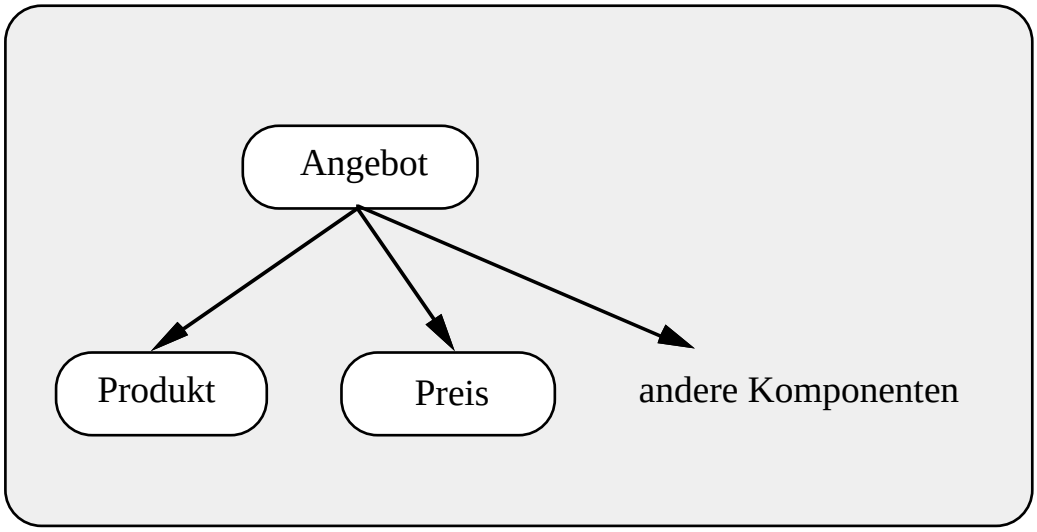


Bild **6.3**: Teil der CSD-Struktur vom Typ "Angebot"

Bei der Untersuchung der 48 Briefe wurde festgestellt, daß zwei Schwierigkeiten auftreten können:

i) Eine klassifikationsrelevante Angebotsphrase kann fehlen.

ii) Es gibt verschiedene Beschreibungsmöglichkeiten für den Artikel.

Prinzipiell können drei verschieden Arten der Beschreibung eines Artikels vorkommen:

a) Der Artikelname wird genannt werden, z.B.:

<artikelname>: "... wir bieten eine M-32 an ..."

b) Die Produktklasse wird benutzt, z.B.:

<produktklasse>: "... wir bieten eine Workstation an ..."

c) Das Modell wird verwandt, z.B.:

<model>: "... wir bieten unsere Workstation M-32 an ..."

und

"... die Ausführung XU-3 kostet ..."

oder "... eine M-32 XL kostet ..."

Aus diesem Grund wird die CSD-Struktur zu einer klassifikationsinternen Struktur erweitert.

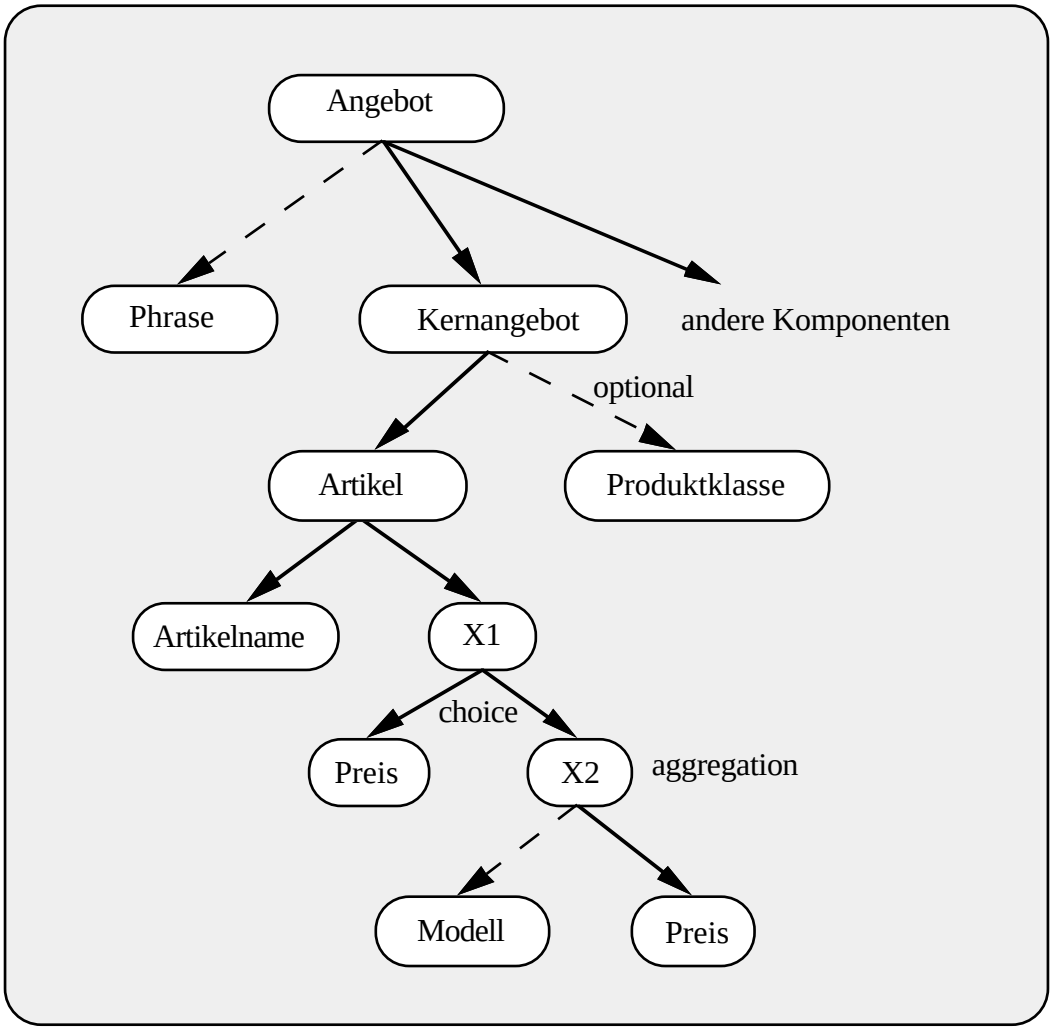


Bild **6.4**: Erweiterte CSD-Struktur, vom Typ "Angebot"

Hierbei werden Komponenten für die terminalen Knoten direkt im Text identifiziert und durch Zeiger miteinander verbunden. Komponenten für komplexe Konzepte werden durch Anwendung von Kombinationsregeln auf die darunterliegenden Komponenten ermittelt. In der Studie der 48 Beispieldokumente wurde untersucht, welche Ausprägungen die Instantiierungen der Komponenten Preis, Produktklasse, Artikel, Modell und Phrase annehmen können. Dabei wurde weiter betrachtet, wie die einzelnen Komponenten im Text auftreten und in Beziehung zu anderen Komponenten stehen sowie die Häufigkeiten ihres Vorkommens in den Beispieldokumenten.

Das anschließende Beispiel listet Phrasen auf, in denen eine Kombination beliebiger Komponenten, zumindestens aber der Preis enthalten ist.

Beispiel:

Häufigkeit	Phraseninstantiierung

2	.. <artikelname>.nom *kostet* <preis>
1	.. <produktklasse>.nom *kostet etwa* ..
1	.. *Preise* <artikelname>.nom *:*<preise> <artikelname>.nom *:*<preis> ..
1	.. *beginnen die Preise bei* <preis> ..
1	.. *der Preis für* <artikelname>.nom *beträgt* <preis> ..
1	.. *Preise liegen zwischen* <preis> *und* <preis> ..
1	.. <artikelname> *hat einen Grundpreis von* <preis> ..
1	.. *gilt der Paketpreis* .. <preis> ..
1	.. *erhalten Sie für* ..
1	.. *bieten* .. <artikelname>.akk .. *zu* .. *von* <preis> .. *zu beziehen* ..
1	.. *bieten* .. *die Möglichkeit* <artikelname>.akk .. *zu beziehen* ..
1	.. *liefern* <produktklasse> ..

.nom = Nominativform vorstehender Komponente
.akk = Akkusativform vorstehender Komponente

Aufgrund der Ergebnisse der Studie wurde folgendes Konzept zur automatischen Generierung von Inhaltszeigern auf die Instantiierungen in den Dokumenttexten vorgestellt.

1) Es werden die Wörter in den Dokumenten erkannt und durch einen Morphologieexperten auf ihre Grundformen gebracht, welche in einem Syntaxlexikon stehen. Nicht-erkannte Token entsprechen dem Muster <unknown>. Inhaltsteile sind Einzelwörter, Mehrwortgruppen und Wortkomplexe.

2) Eine objektorientierte Wissensbasis wird benutzt, um direkt von CDL-Prädikaten auf Elemente innerhalb der Wissensbasis zugreifen zu können. Die folgenden Einträge sind durch einen Zugriff abrufbar:

   a) Instanzen einer Klasse;

   b) Unterklasse einer Klasse;

   c) Komponenten einer Klasse;

   d) Synonymbegriffe.

   Weiterhin sind auch Phrasen als Einträge vorgesehen.

3) Die Erkennung von Standardmustern, wie z.B. Zahlen, Preis- und Datumsangaben, wird von einem Experten vorgenommen. Als <unknown>-Einträge werden nicht identifizierbare Muster gekennzeichnet und weiterbearbeitet. Es wird z.B. ein <unknown>-Eintrag mit den gespeicherten Artikelnamen verglichen und versucht, ihn dadurch zu identifizieren. Ein unbekannter Artikelname soll durch einen Experten ermittelt werden, der mit Regeln wie der folgenden arbeitet:

<artikelname> ::= <produktklasse> directly_before <unknown>.

Um inhaltstragende Teile für komplexe Komponenten zu finden, werden bestimmte Kombinationsbedingungen angewendet, die in [Rieder87] genauer erklärt werden.

7. Ausblick und Erweiterung

Der Einsatz von statistischen Verfahren aus dem Information Retrieval, die in der vorliegenden Diplomarbeit verwendet wurden, haben in den vorliegenden Tests gute Ergebnisse geliefert. Besonders die Bestimmung der Nachrichtentypen für Geschäftsbriefe durch das Klassifikationsmodul wies bei bestimmten Parametereinstellungen eine hohe Akzeptanz der Ergebnisse auf.

Diese Ergebnisse wurden aber nur im Rahmen eines stark eingeschränkten Kontextes erzielt, d.h. unter Benutzung weniger Nachrichtentypen und einer geringen Anzahl von Geschäftsbriefen. Um eine Verbesserung der Ergebnisse zu erreichen, gerade in Hinsicht auf eine größere Anzahl von modellierten Nachrichtentypen, wurden in den entsprechenden Abschnitten Vorschläge und Hinweise auf mögliche Verbesserungen und Erweiterungen aufgezeigt. Die wichtigsten seien im folgenden, auf die jeweiligen Module bezogen, zusammengefaßt:

i) Indexierer

Da die Verfahren aus dem Information Retrieval für große Datenmengen konzipiert wurden, muß die aktuelle Dokumentation unbedingt erweitert werden. Dafür wäre es aber notwendig, die Datenstruktur und die Suchfunktionen der Datenbasis zu überdenken und gegebenenfalls neu zu gestalten, da ansonsten mit einem starken Zeitaufwand zu rechnen ist.

Bisher ist es nur schwer möglich die Ergebnisse der Deskriptorgewichtung durch die einzelnen Gewichtsfunktionen miteinander zu vergleichen, da keine einheitlichen Wertebereiche festgelegt sind. Dieser Nachteil kann durch eine Normierung der Berechnungsergebnisse behoben werden.

Ein weiterer Verbesserungspunkt wäre eine kosequentere Integrierung der Gewichtsfunktion *Diskriminanzwert* [Willett85] oder die Erstellung neuer Gewichtsfunktionen, die noch mehr auf die speziellen Anforderungen der Dokumentenanalyse abgestimmt sind. Hierbei könnten Clusterverfahren ergänzend eingesetzt werden, z.B. Wortcluster für nachrichtentypspezifische Wortlisten.

Weiterhin wäre eine exaktere morphologische Untersuchung des Brieftextes von Vorteil, um bei dem meist geringem Umfang der Texte in Geschäftsbriefen, den Informationsverlust möglichst klein zu halten (Stichwörter: Eindeutigkeit der Stammformengenerierung, Erkennung von Abkürzungen, Bindestrichkomposita, Produktnamen usw.).

Weil nicht mit einer exakten Erkennung aller Buchstaben und Wörter zu rechnen ist, wäre eine Erweiterung der Wortverarbeitung innerhalb des INFOCLAS-Systems um Wortalternativen, evtl. mit Wildcards, wichtig.

ii) Fokussierer

In diesem Modul wäre es sinnvoll, eine flexiblere Gestaltung der Fokusgröße zu ermöglichen, d.h. eine Option zur Angabe eines Bereichs von Deskriptoren zur Bildung des Fokus.

Desweiteren würden Erweiterungen in den Modulen *Indexierer* und *Klassifizierer* sich positiv auf die Berechnungen des Fokussierers auswirken, z.B. neue und verbesserte Gewichtsfunktionen, eine größere Dokumentation oder eine genauere Vorklassifizierung.

Auch ist zu überlegen, ob eine Normierung der Satz- bzw. Deskriptorbereichsgewichte relevant ist und evtl. entfallen sollte.

iii) Klassifizierer

Für die Klassifikation wäre es von Vorteil, wenn die Anzahl der Nachrichtentypen erweitert würde, um die Anzahl der klassifizierbaren Brief zu vergrößern. Dafür ist es aber notwendig, die nachrichtentypspezifischen Wortlisten exakter und umfangreicher aufzustellen. Hilfsmittel dabei könnten Synonymwortlisten und Thesauri sein ([Salton83], [Jüttner88], [Mili&Rada88]).

Ein zusätzliche Erweiterung wäre die Einführung eines Nachrichtentyps, in den Briefe eingeordnet werden, die zu keinem der anderen Nachrichtentypen passen. Beispielsweise könnte man als Ergebnis den Nachrichtentyp *NOCLASS* ausgeben, falls die Ergebnisse der Klassifizierung für die modellierten Nachrichtentypen zu gering sind (Probleme der Modellierung).

Globale Verbesserungen und Erweiterungen, die das gesamte System betreffen, wäre einerseits eine Visualisierungskomponente, die den zu bearbeitenden Brief anzeigt und je nach Modul und Berechnungsergebnis die Deskriptoren hervorhebt oder den Fokus anzeigt. Zweitens könnten als Grundlagen der verwendeten Verfahren nicht mehr nur die einfachen Wortstämme und logischen Objekte fungieren, sondern auch eine Einbeziehung von kontextueller Information (über Wörter hinausgehend) erwogen werden. Zu erwähnen sind in diesem Zusammenhang Ansätze von Salton [Salton91a] [Salton91b] über *globale und lokale Ähnlichkeiten* (*global and local similarities*) oder regelbasierte Analysen bzw. ähnliche Ansätze ([Sacco84], [Lebowitz85], [McCune et al85], [Hayes et al88]).

Wie man sieht, ist die Palette der Verbesserungen und Erweiterungen sehr groß. Auch werden viele Anregungen von ebenfalls auf diesem Gebiet arbeitenden Wissenschaftlern geliefert, so daß noch ein großes Potential in der Verwendung von Information Retrieval Verfahren zur Analyse von Dokumenten existieren.

Literaturhinweise

[Allen87] J. Allen. Natural *Language Understanding*. The Benjamin/Cummings Publishing Company, Inc., Menlo Park, California, 1987.

[Balzert88] H. Balzert. *Wissensbasierte Systeme im Büro der Zukunft*. Symposium "Büro der Zukunft", DECollege, digital, 09.03 - 11.03.88, Stuttgart, 1988.

[Barbic&Rabitti85] W. Barbic, F. Rabitti. *The Type Concept In Office Document Retrieval*. In: Pirotte, A. Vassiliou (eds.): 11th International Conference on Very Large Data Bases, Stockholm, 1985.

[Bleisinger&Hoch&Dengel91] R. Bleisinger, R. Hoch, A. Dengel. *ODA-based modeling for document analysis*. DFKI Technical Memo, TM-91-14, November 1991, 14 pages.

[Bosko78] H. Bosko. *Indexing Concepts and Methods*. Academic Press, 1978.

[Bourne85] S. R. Bourne. *Das UNIX System*. Addison-Wesley, 1985.

[Croft84] B. Croft. *Implementing a text storage and retrieval tool for the office*. IEEE Office Automation Symposium, 1984, pp. 137-143.

[DeJong82] G. DeJong. *An Overview of the FRUMP System*. In W. G. Lehnert, M. H. Ringle (eds.), Strategies for Natural Language Processing. Lawrence Erlbaum Associates, Hillsdale, 1982, pp. 149-175.

[Dengel et al92a] A. Dengel, R. Bleisinger, R. Hoch, F. Fein, F. Hönes, M. Malburg. *Π_{ODA}: The Paper Interface to ODA*. DFKI Technical Report, RR-92-02, February 1992, 53 pages.

[Dengel et al92b] A. Dengel, R. Bleisinger, R. Hoch, F. Fein, F. Hönes. *From Paper to Office Document Standard Representation*. IEEE Computer Magazine, special issue on document image analysis, July 1992.

[Dengel&Hoch92] A. Dengel, R. Hoch. *Intelligent Interfaces between Paper and Computer*. International Symposium on Intelligent Workstations for Professionals, Siemens, München-Neuperlach, March 1992. Will be published in: Lecture Notes, Springer-Verlag, 1992.

[Dittrich92] S. Dittrich. *Analyse von Geschäftsbrieftexten zur Unterstützung der Briefklassifizierung*. Ausarbeitung DFKI Kaiserslautern, 1992.

[Eirund&Kreplin88] H. Eirund, K. Kreplin. *Knowledge Based Document Classification Supporting Integrated Document Handling*, Proc. COIS´88 — Office Information Systems, Palo Alto, CA, March 1988.

[Eirund87] H. Eirund. *Formal Specification of the MULTOS Document Analyser*. Technical Paper, TA Triumph-Adler AG, 1987.

[Eirund88] H. Eirund. *Knowledge Based Document Classification Supporting Content Based Retrieval And Mail Distribution*. Technical Paper, TA Triumph-Adler AG, 1988.

[Finkler86] W. Finkler, G. Neumann. *MORPHIX – Ein hochportabler Lemmatisierungsmodul für das Deutsche*. Universität des Saarlandes (SFB 314), Memo Nr. 8, Saarbrücken, 1986.

[Finkler88] W. Finkler, G. Neumann. *MORPHIX – A Fast Realization of a Classification-Based Approach to Morphology*. Proceedings 4. Österreichische Artificial-Intelligence Tagung. Springer Verlag, Berlin 1988. pp. 11-19.

[Frank91] U. Frank. *Anwendungsnahe Standards der Datenverarbeitung: Anforderungen und Potentiale — Illustriert am Beispiel von ODA/ODIF und EDIFACT* Wirtschaftsinformatik, 33, Jahrgang, Heft 2, April 1991.

[Grochla et al81] E. Grochla et al. *Handbuch der Textverarbeitung*. Verlag Moderne Industrie, Landesberg, 1981.

[Handke89] J. Handke, *Natürliche Sprache: Theorie und Implementierung in LISP*. McGraw-Hill Book Company Gmbh, 1989.

[Hayes et al88] P. J. Hayes, L. E. Knecht, M. J. Cellio. *A News Story Categorization System*. Proc. of Second Conference on Applied Natural Language Processing, 9-12 February, Austin, Texas, 1988.

[Hoch&Malburg92] R. Hoch, M. Malburg. *Designing a Structured Lexicon for Document Image Analysis*. Will be published in: Proc. Seventh Intl. Summer School on Information Technologies and Programming, 28 June - 5. July, Sofia, 1992.

[Horak85] W. Horak. *Office Document Architecture and Office Document Interchange Formats: Current Status of International Standardization*. IEEE Computer Magazine, October 1985, pp. 50–60.

[ISO8613] ISO 8613 Information Processing, Text and Office Systems. *Office Document Architecture and Interchange Format (ODA/ODIF)*, parts 1-8, 1988.

[ISO8879] ISO 8879 Information Processing, Text and Office Systems. *Standard Generalized Markup Language (SGML)*, 1986.

[ISO9735] ISO 9735 *Electronic date interchange for administration, commerce and transport (EDIFACT)*. Application level syntax rules, 1988.

[Jones91] K. Sparck-Jones. *Notes and References on Early Automatic Classification Work*. *SIGIR-Forum*, A Publication of the Special Interest Group on Information Retrieval, Spring 1991, vol. 25, no. 1, pp. 10-18.

[Jüttner88] G. Jüttner, U. Güntzer. *Methoden der Künstlichen Intelligenz für Information Retrieval*. K. G. Saur Verlag, München, 1988.

[Keene89] S.E. Keene. *Object-Oriented Programming in Common LISP*. Addison-Wesley, 1989.

[Kreplin86] K. Kreplin, H. K. Rieder. *EPIKUR*, Forschungsbericht, Wissensbasierte Post-Klassifikationssysteme, 1986.

[Lebowitz85] M. Lebowitz. *Reseacher: An Experimental Intelligent Information System*. Proc. of 9th Intern Joint Conf. on Artificial Intelligence, Los Angeles, Ca, 1985.

[Lutze91] R. Lutze, A. Kohl (Hrsg.). *Wissensbasierte Systeme im Büro, Ergebnisse aus dem WISDOM-Verbundprojekt*. R. Oldenbourg Verlag, 1991.

[Lutz89] E. Lutz, *MULTOS*, TA Triumph-Adler AG Nürnberg, Draft 1.1.

[Manekeller91] W. Manekeller, H. Kirst. *Moderne Korrespondenz*,.Falken Verlag, 1991.

[McCune et al85] B. P. McCune, R. M. Tong, J. S. Dean, D. G. Shapiro. *RUBRIC: A System for Rule-Based Information Retrieval*. IEEE Transactions on Software Engineering, Vol. SE-11, No. 9, September 1985,pp. 939-945.

[Meier78] Helmut Meier. *Deutsche Sprachstatistik*. Georg Olms Verlag, Hildesheim, 2. erweiterte und verbesserte Auflage, Paperback, Band 31, 1978 (1. Auflage 1964).

[Mili&Rada88] H.Mili, R. Rada. *Merging Thesauri: Principles and Evaluation*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 10, No. 2, March 1988.

[Mresse84] M. Mresse. *Information Retrieval — Eine Einführung*. B. G. Teubner Stuttgart, 1984.

[Rieder87] H. G. Rieder. *Einsatz linguistischer Methoden zur Identifikation von Komponenten einer Dokumentstruktur in realen Dokumenten*. WISDOM, Arbeitsbericht, TA Triumph-Adler AG, 1987.

[Sacco84] G. M. Sacco. *OTTER — An Information Retrieval System for Office Automation*. Proc. 2nd ACM SIG OA (Special Interest Group on OA), Conference on Office Automation Systems, Toronto, 1984.

[Sager81] N. Sager. *Natural Language Information Processing*. Addison-Wesley Publishing Company. Reading, Massachusetts, 1981.

[Salton71] G.Salton. *The SMART Retrieval System*. Prentice-Hall, 1971.

[Salton75] G.Salton. *Dynamic Information And Library Processing*. Prentice-Hall, 1975.

[Salton83] G. Salton, M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill Computer Science Series, McGraw-Hill, Inc., 1983.

[Salton86] G. Salton. *Another Look at Text-Retrieval Systems*. Communications of the ACM, vol. 29, no. 7, July 1986.

[Salton87] G.Salton, M.J.McGill. *Information Retrieval — Grundlegendes für Informationswissenschaftler*. McGraw-Hill, 1987.

[Salton89] G.Salton. *Automatic Text Processing*. Addison-Wesley, 1989.

[Salton91a] G. Salton, C. Buckley. *Global Text Matching for Information Retrieval.* Science, vol. 253, August 1991, pp. 1012-1015.

[Salton91b] G. Salton. *Developments in Automatic Text Retrieval.* Science, vol. 253, August 1991, pp. 974-980.

[Steele90] G. L. Steele JR. et al. *Common LISP The Language*, Digital Press, 1990.

[Stürmer91] R. Stürmer. *Modellierung der Brieftexte von Geschäftsbriefen*. Ausarbeitung DFKI, Kaiserslautern, 1991.

[vanRijsbergen79] C.J. van Rijsbergen. *Information Retrieval.* Butterworth, 1979.

[vanRijsbergen84] C. J. van Rijsbergen (ed.). *Research and Development in Information Retrieval.* Proc. of the third joint BCS and ACM symposium King´s College, Cambridge, 2-6 July, 1984, Cambridge University Press, 1984.

[Willett85] P.Willett. *An Algorithm for the Calculation of Exact Term Discrimination Values*. Pergamon Press Ltd., Information Processing & Management, Vol. 21, No. 3, pp. 225-232. 1985.

[Yu et al82] C. T. Yu, K. Lam, G. Salton. *Term Weighting in Information Retrieval Using the Term Precision Model.* Journal of the Association for Computing Machinery, Vol. 29, No. 1, January 1982, pp. 152-170.

Anhang

A. Briefnummern und zugehörige Nachrichtentypen (manuelle Klassifizierung)

In der folgenden Aufstellung wird die manuelle Klassifizierung der vorhandenen Geschäftsbriefe angegeben. In der ersten Spalte steht dabei die Briefnummer, in der zweiten die Abkürzung des Nachrichtentyps. In Sonderfällen können auch zwei Nachrichtentypen einem Brief zugeordnet sein. Weiterhin treten zwei Nachrichtentypen auf, die nicht in INFOCLAS modelliert sind (DATN Datenübertragung, BEND Bestelländerung).

Brief1	(BEBE WERB)	Brief43	(BEBE)
Brief2	(BEBE)	Brief44	(BEBE)
Brief3	(BEBE WERB)	Brief47	(BEBE)
Brief4	(WERB)	Brief48	(BEBE)
Brief5	(BEST)	Brief49	(ANFR)
Brief6	(BEBE)	Brief50	(BEBE)
Brief7	(BEBE)	Brief51	(BEBE)
Brief8	(ANFR)	Brief54	(BEBE)
Brief9	(BEBE)	Brief55	(DATN)
Brief10	(BEBE)	Brief56	(DATN)
Brief11	(BEBE ANGE)	Brief57	(BEBE)
Brief12	(BEBE)	Brief59	(BEBE)
Brief13	(BEST)	Brief60	(BEBE)
Brief14	(BEBE ANFR)	Brief62	(BEBE)
Brief15	(BEBE ANFR)	Brief63	-
Brief16	(WERB)	Brief64	-
Brief17	(ANFR)	Brief65	(BEST)
Brief18	(BEBE)	Brief66	(ANGE)
Brief19	(BEST)	Brief67	-
Brief20	(BEBE)	Brief68	(DATN)
Brief21	(BEBE)	Brief69	(ANGE)
Brief22	(DATN)	Brief70	(BEST)
Brief23	(ANGE DATN)	Brief71	(BEST)
Brief24	(DATN)	Brief72	-
Brief25	(ANFR)	Brief74	(BEST)
Brief26	(ANFR)	Brief75	(DATN WERB)
Brief27	(ANFR)	Brief76	(BEBE BEST)
Brief28	(ANFR)	Brief77	(BEST)
Brief29	(ANFR)	Brief78	-
Brief30	(DATN WERB)	Brief79	(ANGE)
Brief32	(DATN)	Brief80	-
Brief33	(DATN BEBE)	Brief81	(BEST)
Brief34	(BEND)	Brief82	(BEST)
Brief35	(BEBE)	Brief83	(BEBE ANFR)
Brief36	(BEBE)	Brief84	-
Brief37	(BEBE)	Brief85	-
Brief38	(DATN)	Brief86	-
Brief39	(ANFR)	Brief87	-
Brief40	(ANFR)	Brief88	-
Brief41	(ANFR)	Brief89	-
Brief42	(BEBE)		

B. Nachrichtentypspezifische Wortlisten

In diesem Abschnitt werden die spezifischen Wortlisten der einzelnen Nachrichtentypen aufgelistet. Die Einträge in den Listen sind Wortstämme, die von MORPHIX ermittelt wurden. Für jeden Nachrichtentyp wurde eine Einteilung der Wortlisten in Primärwort-, Sekundärwort- und Tertiärwortlisten vorgenommen. Exemplarisch wurden die Nachrichtentypen *Anfrage* und *Werbung* ausgewählt.

1) ANFRAGE

Primärwortliste:

("bitt" "frag" "moeglich" "moeglichkeit" "moeglichst" "schick" "send" "zusendung" "uebersend" "zusend" "nachsendung" "uebersendung")

Sekundärwortliste:

("dank" "arbeit" "schreib" "wichtig" "autor" "neu" "wuerde" "erhalt" "artikel" "dankbar" "interessier" "lass" "veroeffentlichung" "zukomm" "aehnlich" "anlage" "bedank" "direkt" "einzeln" "entgegenkomm" "erstell" "exemplar" "kopie" "nenn" "termin" "versand" "ansichtsexemplar" "ansprechpartner" "antwort" "anzahl" "ausgabe" "ausleih" "auszug" "baldig" "beifueg" "beiliegend" "bekomm" "beschreib" "beschreibung" "freu" "geb" "hoeflich" "informier" "interessant" "kostenangebot" "kostenlos" "umfang" "verfuegung" "zugaenglich" "forder")

Tertiärwortliste:

("herr" "gruss" "werd" "ehr" "freundlich" "beitrag" "studie" "konferenz" "projekt" "viel" "gross" "information" "institut" "bibliothek" "computerwoche" "fall" "informatik" "kompaktseminar" "kuenstlich" "reis" "akademisch" "all" "anfang" "belegheft" "bereich" "expertensystem" "fortdruck" "gesellschaft" "intelligenz" "landesforschungsbericht" "nah" "oesterreichisch" "system" "uniservice" "wissensingenieur" "abzug" "arbeitsgruppe" "arbeitsmarkt" "beabsichtig" "bedeutung" "beid" "berufsbild" "berufsleben" "biet" "deutsch" "end" "folg" "folie" "forschung" "forschungslaboratorium" "frau" "institution" "jahr" "juni" "kontakt" "korrigier" "kurz" "maurer" "mensch" "nachwuchs" "name" "oktober" "personal" "plan" "praesentation" "professor" "prospekt" "prozessleittechnik" "recht" "seite" "setz" "spei" "team" "teilnehm" "these" "universitaet" "unterstuetz" "verbleib" "verlag" "verteil" "vorstell" "zwischenzeit" "ablichtung" "abseh" "absolvent" "abstimmung" "abteilung" "allgemeinverstaendlich" "amerikanisch" "analysis" "ander" "ankuendig" "annahme" "annehm" "anwender" "april" "arbeitslos" "arbeitsueberlastung" "argumentationshilfe" "aspekt" "auflage" "aufnehm" "augustwoche" "ausbild" "ausseh" "ausserordentlich" "band" "beauftragt" "bedarf" "befragung" "befuerwort" "beginn" "bemerk" "benutz" "beratungsunternehmen" "beruf" "berufsanfaenger" "besitz" "besprechung" "beteilig" "bezug" "blick" "brems" "brief" "briefbreite" "bundesanstalt" "dame" "darstell" "darstellung" "datenverarbeitung" "datum" "direktor" "dokumentwissen" "duden" "effizient" "eign" "einhaltung" "einig" "einlad" "einladung" "einplan" "einreich" "einsatz" "einsatzrisiko" "einstiegshilfe" "einzeilig" "endgueltig" "entscheid" "entscheidung" "entwickel" "entwicklung" "erfahr" "erfahrungswissen" "erforderlich" "ergebnis" "erlaeuter" "eroerter" "erschein" "erstmalig" "erwaehn" "erwartung" "erzaehl" "experte" "expertensystementwicklung" "fachlich" "fakultaet" "falsch" "familie" "finanzierung" "firma" "flugfuehrung" "foerderlich" "foerderung" "formblatt" "forschungsaktivitaet" "forschungsbericht" "forschungseinrichtung" "forschungsinstitut" "forschungsprojekt" "forschungszentrum" "foto" "frisch" "gehoer" "gelegenheit" "gemeinsam" "gesamtgesellschaftlich" "gespraechstermin" "gut" "hamburger" "hauptprojekt" "herausgeb" "hersteller" "herzlich" "hilfe" "hochschule" "informationstechnik" "informationstechnisch" "ingenieurwissenschaftlich" "inhaltsverzeichnis" "intelligent" "irrtum" "jaehrlich" "kalenderwoche" "kapitel" "kennzeichn" "klein" "knuepf" "koautor" "kommentar" "kommunikationsbranche" "kompass" "komponente" "konferenzband" "konzept" "korrekturfahne" "kultusministerium" "kuratorium" "kurzfristig" "kurzgliederung" "kurzinformation" "labor" "land" "lauf" "layout" "leiter" "letzt" "lieb" "liste" "maerz" "markt" "marktforschungsinstitut" "maschine" "mathematik" "meinung" "meld" "menge" "mitarbeit" "mitarbeiter" "mitglied" "mitte" "mitwirk" "mitwirkung" "nachdruecklich" "nachricht" "nachsteh" "natuerlich" "nehm" "nichtuniversitaer" "noetig" "normal" "oeffentlich" "oesterreich" "offen" "offenbar" "organisation" "orientierung" "pass" "potentiell" "praxis" "programmiersprache" "prolog" "publikation" "qualifikation" "rahm" "redaktionsschluss" "reklamier" "restriktion" "richt" "rolle" "schluesseltechnologie" "schreibmaschine" "schreibmaschinenseite" "seitig" "selbstverstaendlich" "sicherheitsempfindlich" "sitzung" "software" "softwarehaus" "sonder" "sonderdruck" "sonderpublikation" "sonstig" "speziell" "sponsor" "sprachlich" "sprech" "stell" "steuer" "student" "studienrichtung" "szene" "taetigkeitsfeld" "tageszeitung" "tagungsband" "tagungsbericht" "technologie" "technologiezentrum" "telefonnummer" "terminlich" "thema"

"titel" "transfer" "treff" "trend" "uebernehm" "ueberschreit" "umfass" "umgeschult" "unterlauf" "unterlieg" "unternehm" "unterricht" "unterstell" "veraender" "verantwortlich" "verantwortung" "verbindung" "vereinbar" "verfasser" "vergess" "vergleich" "verhandlungsfuehrend" "verlier" "vermehr" "veroeffentlich" "verschieden" "versprech" "verstaendigung" "vertret" "vision" "vorfahr" "vorhaben" "vorhanden" "vorleg" "vorprojekt" "vorseh" "weiss" "weitgehend" "wissensbasiert" "wissensbasis" "wissenschaftlich" "wissenschaftsminister" "wolf" "zeilig" "zeit" "zentrum" "zugreif" "zusaetzlich" "zusammenarbeit" "zusammenfass" "zusammenhang" "zustaendig")

2) WERBUNG

Primärwortliste:

("neu" "information" "biet" "kunde" "ueberzeug" "kostenangebot" "preisvorteil" "werbebrief" "werbeprospekt" "werbung")

Sekundärwortliste:

("freundlich" "vorteil" "aktuell" "ausgabe" "frag" "geb" "erfolg" "gut" "moeglichkeit" "preis" "produktion" "produzier" "ansprechpartnerin" "antwort" "bekomm" "besonder" "bestmoeglich" "besuch" "brauch" "dank" "einfach" "einzigartig" "enthalt" "entscheidungshilfe" "erfahr" "erforderlich" "ergebnis" "erhalt" "erschein" "erwaehn" "erwart" "garantie" "garantier" "geld" "gross" "gruendlich" "hervorragend" "hoch" "idee" "info" "informationsmaterial" "kennenlern" "klar" "kompetent" "konkurrenz" "kundenfreundlich" "leicht" "liefer" "lieferbar" "markt" "nachdenk" "natuerlich" "optimier" "persoenlich" "praesentier" "produkt" "qualitaet" "realistisch" "reservier" "selbstverstaendlich" "service" "servicenummer" "sonderpreis" "uebersend" "umfang" "verbesser" "verfueg" "verfuegung" "verkauf" "vorlieg" "vorschlag" "wesentlich" "wuensch" "zusammenarbeit" "zusend" "zutreff")

Tertiärwortliste:

("lufthansa" "gruss" "impuls" "werd" "ehr" "video" "bitt" "herr" "servicekarte" "tu" "denk" "gehoer" "gesellschaft" "gruendung" "lieb" "monat" "projekt" "wissen" "ander" "betreib" "bewertung" "deutsch" "erreich" "fax" "film" "find" "folg" "freu" "hand" "medium" "microvax" "nenn" "professor" "rechner" "sicher" "software" "steh" "studie" "system" "teilsystem" "uebersicht" "unterstuetzung" "verlag" "vorbereit" "vorhanden" "weit" "wirtschaftsmagazin" "wort" "abonnement" "abstuetz" "aktiengesellschaft" "angebotsfrage" "animation" "anruf" "anschrift" "anspruch" "arbeit" "architektur" "aufgabe" "auflage" "auftrag" "ausdruecklich" "begrenz" "beileg" "bekannt" "benoetig" "benutzerschnittstelle" "bereichsleiter" "beschreib" "besitz" "betracht" "bild" "bombardier" "bueroraum" "cd" "cpu" "dame" "darstell" "darstellungsmoeglichkeit" "datum" "decstation" "digital" "diskutier" "druckfrisch" "eigen" "eigentlich" "eign" "einfuehr" "eingabe" "einrichtung" "einsetz" "einteil" "einzeln" "einzig" "empfang" "end" "entscheidung" "entscheidungstraeger" "entsprech" "entwickel" "ergebnisbericht" "experte" "fachgebiet" "fachlich" "faxgeraet" "fernsehstandart" "finanz" "finqer" "firma" "flugfuehrung" "formel" "gedanke" "geh" "geldwert" "geraet" "gesamt" "geschaeft" "geschaeftsbedingung" "gestalt" "gleich" "gleichzeitig" "gratisexemplar" "grund" "grundsaetzlich" "kapazitaetsberechnung" "karte" "koepfig" "kommunikationsinterface" "komponente" "komponentenentwicklung" "kuendig" "kuenftig" "kundenstamm" "lager" "laut" "les" "leser" "leserin" "lieg" "luftverkehr" "mai" "massenkopierung" "mathematisch" "merk" "mess" "nachricht" "name" "nehm" "norm" "normwandlung" "nutzbar" "originalheft" "papierkorb" "partner" "peripherie" "personalaufstockung" "plan" "platte" "plotter" "postleitzahlbereich" "potentiell" "praesentation" "printwerbung" "problem" "projektleiter" "realisierung" "rechnergestuetzt" "rechnungswesen" "recht" "redaktion" "regelmaessig" "regelung" "ruf" "ruh" "sag" "scheu" "schnittstelle" "seite" "sendefaehig" "setz" "sicherstell" "skeptisch" "speicherung" "sprache" "steuer" "stichwort" "streich" "teilgebiet" "telefax" "telefonverkauf" "tisch" "trickmischung" "uebertrag" "uebersicht" "umsetz" "umzug" "unlimitiert" "unternehm" "untersuch" "urlaubsfilm" "vax" "vaxstation" "verantwortungsbereich" "verreis" "versorg" "versteh" "versuchsprojekt" "versuchssystem" "videopraesentation" "videothek" "visuell" "voraussetzung" "welt" "werf" "west" "wissensspeicherung" "wohn" "wunder" "zeit" "zeitgenosse" "zoll" "zulauf")

Index