Universität Leipzig
Fakultät für Mathematik und Informatik
Institut für Informatik

Text, Wörter, Morpheme — Möglichkeiten einer automatischen Terminologie-Extraktion

Diplomarbeit

Leipzig, März 2004

vorgelegt von:
Hans Friedrich Witschel

Vorwort

Diese Arbeit ist in Kooperation mit der Firma GlobalWare AG entstanden. Ich möchte mich für die gute Zusammenarbeit bedanken, die mir viele Einblicke in die Terminologiearbeit im Rahmen der maschinellen Übersetzung ermöglichte und somit mein Verständnis für den Gebrauch und die Notwendigkeit der Terminologie-Extraktion förderte.

Mein besonderer Dank gilt Herrn Dr. Frank Beckmann, der mir den Zugang zu dem Übersetzungssystem LOGOS ermöglichte und mir während der gesamten Arbeit mit Rat und Tat zur Seite stand.

Inhaltsverzeichnis

Einleitung

Die Suche nach Informationen in unstrukturierten natürlichsprachlichen Daten ist Gegenstand des sogenannten *Text Mining*.

In dieser Arbeit soll ein Teilgebiet des Text Mining beleuchtet werden, nämlich die Extraktion domänenspezifischer Fachbegriffe aus Fachtexten der jeweiligen Domäne.

Ein Beispiel: aus folgender Passage eines medizinischen Fachtextes:

> The major risks of long-term **cardiotoxicity** relate to treatment prior to the **BMT**, in particular, **anthracyclines**, **ablative-dose** Cytoxan (ie, dose >150 mg/kg), chest **radiation therapy**, **TBI** (especially if not fractionated), and high-dose **steroids**.

sollten mindestens die **fett** gedruckten Termini extrahiert werden. Schon hier zeigt sich aber, daß nicht immer ganz klar ist, welche Wörter zur Terminologie einer Domäne gehören: „Cytoxan“ kann als Name eines Medikamentes extrahiert werden, es gibt aber auch Gründe, die dagegen sprechen.

Es soll also die Frage geklärt werden, was Terminologie eigentlich ist, vor allem aber werden verschiedene Methoden entwickelt, welche die Eigenschaften von Fachtermini ausnutzen, um diese aufzufinden. Die Verfahren sollen aus den linguistischen und „statistischen“ Charakteristika von Fachbegriffen hergeleitet werden und auf geeignete Weise kombiniert werden. Dabei werden sie auch bezüglich ihrer Leistungsfähigkeit verglichen.

Anliegen dieser Abhandlung ist es also, die Möglichkeiten aufzuzeigen, die eine Kombination verschiedener musterbasierter und statistischer Verfahren im Hinblick auf Problemstellungen des Text Mining bietet. Es werden aber auch die Grenzen und Probleme dieser Ansätze angedeutet.

Motivation

Wofür überhaupt Terminologie-Extraktion? Die Antwort darauf ist einfach: der Schlüssel zum Verständnis vieler Fachgebiete liegt in der Kenntnis der zugehörigen Terminologie. Natürlich genügt es nicht, nur eine Liste der Fachtermini einer Domäne zu kennen, um diese zu durchdringen.

Eine solche Liste ist aber eine wichtige Voraussetzung für die Erstellung von Fachwörterbüchern (man denke z.B. an Nachschlagewerke wie das klinische Wörterbuch „Pschyrembel“): zunächst muß geklärt werden, *welche* Begriffe in das Wörterbuch aufgenommen werden sollen, bevor man sich Gedanken um die genaue Definition der einzelnen Termini machen kann.

Ein Fachwörterbuch sollte genau diejenigen Begriffe einer Domäne beinhalten, welche Gegenstand der Forschung in diesem Gebiet sind oder waren. Was liegt also näher, als entsprechende Fachliteratur zu betrachten und das darin enthaltene Wissen in Form von Fachtermini zu extrahieren?

Darüberhinaus sind weitere Anwendungen der Terminologie-Extraktion denkbar:

Maschinelle Übersetzung: Bei der automatischen Übersetzung von Texten wird ein möglichst umfangreiches elektronisches Wörterbuch benötigt. Dieses enthält zu Beginn meist nur wenige fachspezifische Begriffe. Diese können mittels Terminologie-Extraktion identifiziert und eingegeben werden.

Beschlagwortung: Das aus dem Information Retrieval stammende Problem des automatischen Indexierens von Texten ist der Terminologie-Extraktion verwandt: die Liste der Fachtermini eines Textes ist oft ein guter Indikator für dessen Inhalt.

Die Normung von Terminologie, d.h. die Festlegung standardisierter Terminologien ist außerdem ein wichtiger Faktor für die Vereinheitlichung und damit Vereinfachung der Kommunikation zwischen Experten: wenn alle dieselbe Sprache sprechen, versteht man sich besser. Terminologie-Extraktion kann helfen, uneinheitliche Terminologien zu identifizieren und zu vereinheitlichen.

Aufbau der Arbeit

Die Arbeit gliedert sich wie folgt:

Im ersten Kapitel wird versucht, den Begriff „Terminologie“ zu definieren und einzugrenzen. Darüberhinaus möchte ich Eigenschaften von Fachtermini finden, die sich für deren Extraktion ausnutzen lassen. Aus diesen Charakteristika sollen direkt Verfahren bzw. Heuristiken abgeleitet werden, die dann später in ein Gesamtsystem einfließen werden.

Kapitel zwei beschäftigt sich mit bereits vorhandenen Ansätzen zur Terminologie-Extraktion: eine Auswahl der für mich wichtigsten Verfahren wird vorgestellt und bewertet. Es wird der Versuch unternommen, aus den Stärken und Schwächen der vorhandenen Systeme zu lernen und sie im Hinblick auf eigene Ideen zu verwerten.

Im dritten Kapitel wird ein von mir entwickeltes und implementiertes Verfahren beschrieben. Dieses integriert viele der in Kapitel 1 und 2 abgeleiteten Methoden und versucht darüberhinaus, aus dem Feedback des Anwenders zu lernen und die Suche nach Fachtermini iterativ zu verfeinern.

Das vierte Kapitel untersucht die Ergebnisse, die sich mit dem in Kapitel 3 beschriebenen Verfahren erzielen lassen. Dabei sollen dessen Komponenten getrennt untersucht und bezüglich ihrer Leistungsfähigkeit verglichen werden.

Kapitel 5 beschreibt kurz die Veränderungen, welche nötig sind, um das System an andere Sprachen anzupassen. Dies wird am Beispiel der deutschen Sprache durchgeführt; das ursprüngliche Verfahren wurde für Englisch entwickelt.

Schließlich beschäftigt sich Kapitel 6 mit einem Ausblick auf Verfeinerungen und Erweiterungen des Systems, die im Rahmen dieser Diplomarbeit nicht mehr durchgeführt werden konnten.

Kapitel 1

Terminologie — Eingrenzung des Begriffs

1.1 Fachsprachen und Terminologie

Der Begriff „Fachsprache“ wird durch die Deutsche Industrie-Norm (DIN) wie folgt definiert:

> „Bereich der Sprache, der auf eindeutige und widerspruchsfreie Kommunikation in einem Fachgebiet gerichtet ist und dessen Funktionieren durch eine festgelegte Terminologie entscheidend unterstützt wird.“ ([DIN 2342])

Oft wird zusätzlich zu dieser Definition eine Abgrenzung der Fachsprachen von der Gemeinsprache vorgenommen (s.u.), um besser erfassen zu können, wodurch sich Fachsprachen auszeichnen.

Zunächst aber wollen wir uns auf einige Aspekte obiger Definition konzentrieren: Erst einmal ist zu erklären, was wir unter „Terminologie“ verstehen. Eugen Wüster, der Begründer der „Allgemeinen Terminologielehre“, versteht unter Terminologie

> „das Begriffs- und Benennungssystem eines Fachgebietes, das alle Fachausdrücke umfaßt, die allgemein üblich sind.“ ([Wüster 1991], S. V)

Diese Definition hat sich allgemein durchgesetzt und hat auch Eingang in den DIN-Standard gefunden (siehe Tabelle A.2 der DIN-Definitionen im Anhang).

Die Kombination aus Begriff und Benennung hat Tradition: Schon Ferdinand de Saussure, der Begründer des Strukturalismus, unterschied bei sprachlichen Zeichen (also Wörtern) zwischen Zeicheninhalt und Zeichenform, d.h. zwischen abstrakten mentalen Konzepten von Gegenständen und der Art ihrer konkreten sprachlichen Realisierung (vgl. [Saussure 1967], S. 77 f.). Wir erkennen z.B. viele verschiedene Konkreta als Baum, weil wir ein *mentales Konzept* davon haben, was Eigenschaften eines Baumes sind.

Ein „Begriff“ entspricht also dem Zeicheninhalt, d.h. der Bedeutung eines Wortes (in unseren Köpfen); die zugehörige „Benennung“ ist die Zeichenform, d.h. die sprachliche Realisierung. Beide zusammen ergeben dann (laut [DIN 2342]) einen Terminus. Zu unterscheiden ist nach [DIN 2342] noch zwischen Einwort- und Mehrwortbenennungen, was uns später näher interessieren wird.

Was unterscheidet nun Fachtermini von Wörtern der Gemeinsprache? Wie oben in der DIN-Definition für Fachsprache angedeutet, geht es den Terminologen um Eindeutigkeit („eindeutig und widerspruchsfrei“). Wüster betont zunächst, daß einer Benennung eine Sach- bzw. Begriffsbedeutung zugeordnet wird (vgl. [Wüster 1991], S. 2), daß die Mitbedeutungen aber wegfallen.

In anderen Quellen wird oft darauf hingewiesen, daß die Zuordnung zwischen Begriff und Benennung eineindeutig ist bzw. sein soll (vgl. z.B. [Maynard 1999], S. 24 f. oder [Bourigault 1992]). Man spricht daher auch von „Eindeutigkeit“ und „Exaktheit“ (vgl. [Blank 1997], S. 7 f.). Das bedeutet, daß sowohl die Mehrdeutigkeit von Fachtermini ausgeschlossen ist, als auch die Existenz von Synonymen. [Maynard 1999] relativiert dies aber, indem sie feststellt, daß beide Phänomene auch in Fachsprachen auftreten, aber eben seltener als in der Gemeinsprache (vgl. auch [Hoffmann 1988], S. 66).

Wüster spricht auch von „Ist-Normen“ und „Soll-Normen“ und konzentriert sich auf die Erstellung von Soll-Normen zur Auflösung des „untragbaren Durcheinanders“ ([Wüster 1991], S. 3) der vorhandenen Uneindeutigkei-

ten.

Auch [Fluck 1985] sieht die Aufgabe der Fachsprachen darin, einen Zeichenvorrat bereitzustellen, der uns in die Lage versetzt, uns möglichst präzise und ökonomisch über einen bestimmten Sachbereich auszutauschen. Je eindeutiger die Termini, desto größer ist die Präzision, je weniger Synonyme, desto ökonomischer wird die Kommunikation.

Ein weiterer wichtiger Vorteil dieser Eindeutigkeit liegt auch darin, daß sich (normierte) Fachtermini leicht in andere Sprachen übersetzen lassen: Da Begriffe weitgehend sprachunabhängig sind, d.h. die Konzepte, die sich hinter Benennungen verbergen, in allen Sprachen gleich sind (bis auf eventuelle kulturelle Besonderheiten), muß zu einer Benennung einer Quellsprache nur die Benennung der Zielsprache gefunden werden, die dem gleichen Begriff entspricht.

Schließlich sind einheitliche Benennungen zunehmend auch für große Unternehmen wichtig, um — insbesondere bei den Kunden — Verwirrung vorzubeugen.

Neben der angestrebten Eindeutigkeit in der Beziehung zwischen Begriffen und Benennungen beschäftigt sich die Allgemeine Terminologielehre auch mit der genormten Erfassung dieser Beziehungen mittels Definitionen, sowie mit der Anordnung der Termini in Hierarchien.

Die Frage der Normierung spielt also eine wichtige Rolle in der Arbeit mit Terminologien, soll uns im Folgenden aber weniger beschäftigen. Vielmehr möchte ich versuchen, Eigenschaften von Fachtermen herauszuarbeiten, die es erlauben, diese in Fachtexten aufzufinden: die Identifizierung der vorhandenen Terminologien einer Domäne (d.h. der Ist-Normen) ist ein erster wichtiger Schritt auf dem Weg zur Vereinheitlichung.

Dazu werde ich zunächst näher darauf eingehen, wie sich Fachsprachen von der Gemeinsprache unterscheiden.

1.2 Abgrenzung der Fachsprachen von der Gemeinsprache

Die Gemeinsprache wird durch [DIN 2342] als „Kernbereich der Sprache, an dem alle Mitglieder einer Sprachgemeinschaft teilhaben“ definiert. Die Fachsprachen unterscheiden sich von ihr in zwei wichtigen Punkten:

- Im Wortschatz: in der Menge der Fachtermini bzw. Terminologie
- In der Syntax: gewisse syntaktische Konstruktionen der Gemeinsprache werden in Fachsprachen bevorzugt, d.h. mit höherer Frequenz benutzt als andere (z.B. Passivkonstruktionen in technischen Anleitungen).

Da wir mehr an den lexikalischen Besonderheiten der Fachsprachen interessiert sind, wird darauf im Folgenden näher eingegangen:

Der Wortschatz der Fachsprachen läßt sich laut [Fluck 1985] (S. 16-23) einmal horizontal in verschiedene Fachbereiche (Pharmazie, Anatomie etc.) unterteilen, zum anderen vertikal, d.h. nach Sprachebene bzw. Spezialisierungsgrad. Hier gibt es verschiedene Ansätze, vielen gemeinsam ist jedoch die Verwendung folgender Kriterien:

Normung: Begriffe, deren Bedeutung per Definition bzw. Standard festgelegt ist, erscheinen als „höchste“ Form der Terminologie.

Eindeutigkeit: je spezieller ein Fachterminus, desto eindeutiger ist er.

Theorie vs. Praxis: Termini des theoretisch-wissenschaftlichen Stils treten nur in geschriebenen Texten auf, während Fachwörter der sog. fachlichen Umgangssprache der Verständigung unter Fachleuten dienen, also vorwiegend in gesprochener Sprache vorkommen.

Probleme

Es stellt sich nun also die Frage, wann ein Wort als fachsprachlich (d.h. als Fachterminus) und wann als gemeinsprachlich anzusehen ist. Hierauf kann man in vielen Fällen keine eindeutige Antwort geben, vielmehr handelt es

sich um eine graduelle Entscheidung, d.h. manche Wörter sind „fachsprachlicher“ als andere.

Im Hinblick auf die Evaluierung eines Verfahrens zur Terminologie-Extraktion ist es aber sehr wichtig, eine genaue Abgrenzung vorzunehmen: wie soll ich die Qualität eines Ergebnisses beurteilen, wenn ich nicht vorher genau festgelegt habe, was ein „gutes“ Ergebnis ist? Oder anders gesagt: Man sollte erst genau wissen, wonach man sucht, bevor man sich an die Implementierung eines Extraktionssystems macht.

Das Problem der genauen Abgrenzung ist aber kein vollkommen neues: im Information Retrieval sucht man nach „relevanten“ Dokumenten, ohne eine allgemeingültige Definition für Relevanz zu haben. Vielmehr hängt das erwünschte Ergebnis von den Bedürfnissen des Benutzers ab: jeder Benutzer hält etwas anderes für relevant oder — in unserem Fall — für fachsprachlich. Herauskommen soll also genau das, was der Benutzer gerne möchte. Wie man das zumindest ansatzweise garantieren kann, soll in Abschnitt 1.5.2 diskutiert werden.

Trotzdem möchte ich vorher versuchen, ein paar universelle Eigenschaften von Fachtermini herauszuarbeiten, um Ideen für ein Grundgerüst eines Extraktionsverfahrens zu entwickeln.

Zwei wichtige Eigenschaften von Termini lassen sich für ihre Identifikation in Texten ausnutzen: sie folgen einer anderen statistischen Verteilung als gemeinsprachliche Wörter und sie werden nach bestimmten Mustern gebildet, die immer wieder — teilweise auch sprachübergreifend — auftauchen und die sich linguistisch auf bestimmte Weise charakterisieren lassen. Darauf soll im Folgenden näher eingegangen werden.

1.3 Statistische Verteilung von Termini in Fachtexten

Eine Abhandlung über die statistische Verteilung von Wörtern in Texten wird nie ohne die Erwähnung des sogenannten Zipfschen Gesetzes auskommen. Das Zipfsche Gesetz macht eine allgemeine Aussage, die auch in Fachtexten gilt: Ordnet man die Wörter eines beliebigen Textes absteigend nach

ihrer Häufigkeit f und bezeichnet der „Rang“ r eines Wortes seinen Platz in der geordneten Liste, so gilt:

$$f * r = const. \tag{1.1}$$

Daraus lassen sich viele interessante Folgerungen ableiten, z.B. die Tatsache, daß es in jedem Text ein paar wenige Wörter gibt, die sehr häufig vorkommen und somit einen Großteil des Textes ausmachen (die häufigsten 100 Wörter machen etwa die Hälfte eines jeden Textes aus, vgl. [Manning 2002] S. 22 ff.).

Dabei stellt man fest, daß diese extrem häufigen Wörter allesamt rein syntaktische Funktion haben: es handelt sich um Artikel, Präpositionen und Konjunktionen. Diese sogenannten Synsemantika bezeichnen nichts Außersprachliches, sondern erhalten nur im Zusammenhang mit den Inhaltswörtern (Autosemantika) ihres Kontexts einen Sinn. Wir wollen diese sehr häufig auftretenden Synsemantika im Folgenden *Stopwörter* nennen.

In [Hoffmann 1988] findet man eine Untersuchung, welche eine Aufstellung der häufigsten Wörter einiger Fachtexte mit den häufigsten Wörtern allgemeinsprachlicher Texte vergleicht ([Hoffmann 1988], S. 62 ff.). Zusammenfassend lassen sich die Ergebnisse folgendermaßen darstellen:

- Die Stopwörter treten in allen betrachteten Texten mit gleichen relativen Häufigkeiten auf, sind also sowohl in Fachsprachen als auch in der Gemeinsprache unter den häufigsten Wörtern.

- Bestimmte Autosemantika, vor allem Nomina, die in einer bestimmten Fachsprache sehr häufig sind, treten in der Gemeinsprache wesentlich seltener (oder gar nicht) auf. Es drängt sich der Verdacht auf, daß es sich hier eben um die Terminologie der Fachsprache handelt.

- Diese Mengen von Autosemantika sind charakteristisch für spezielle Fachgebiete, d.h. die Mengen der häufigsten Autosemantika zweier Fachdomänen sind weitgehend disjunkt.

- Die Mehrdeutigkeit der fachsprachlichen Autosemantika ist gegenüber der Gemeinsprache stark reduziert, was ein weiterer Hinweis darauf ist, daß sie Fachtermini sind. Um ein Beispiel zu nennen: die Bezeichnung

„Wurzel“ wird in einem mathematischen Text mit hoher Wahrscheinlichkeit nur die Bedeutung einer speziellen mathematischen Operation haben, in der Gemeinsprache kann es sich genausogut um die Wurzel eines Baumes oder Zahnes handeln.

Das bisher Genannte soll durch Abb. 1.1 noch einmal graphisch veranschaulicht werden.

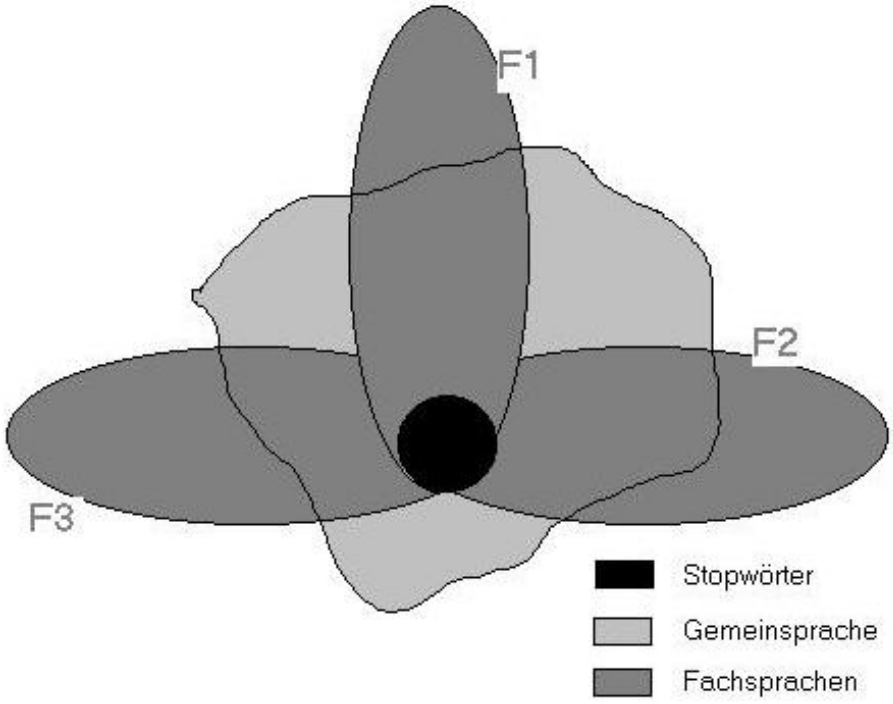

Abbildung 1.1: Verteilung lexikalischer Elemente in Gemeinsprache und Fachsprachen

Ableitung einer Heuristik:

Aus den bisher gewonnenen Erkenntnissen will ich nun eine Heuristik ableiten, um Termini aus Fachtexten zu extrahieren: Dazu betrachte ich zu jedem Wort zwei Parameter:

1. eine geeignete statistische Prüfgröße P, welche die relative Häufigkeit des Wortes im vorliegenden Fachtext vergleicht mit seiner relativen

Häufigkeit in einem allgemeinsprachlichen Korpus. P soll dabei groß werden, wenn ein Wort im Fachtext wesentlich häufiger auftritt, als dies nach seiner Häufigkeit im Korpus zu erwarten gewesen wäre: Man mißt die „Überraschung“, das Wort hier so häufig anzutreffen; sowie

2. die absolute Frequenz des Wortes im aktuellen Fachtext.

Für die Berechnung der Prüfgröße P stellt sich noch die Frage nach einem geeigneten Vergleichskorpus. Diese soll zunächst hintangestellt werden. Wir werden darauf zurückkommen, wenn es an die Implementierung eines Verfahrens geht. Dasselbe gilt für die konkrete Berechnung von P.

Nun also zu der Frage, wie sich mittels dieser beiden Parameter Fachterminologie identifizieren läßt: Wir haben gesehen, daß fachsprachliche Autosemantika (die wir als Fachtermini ansehen wollen) in Fachtexten wesentlich häufiger auftreten als in gemeinsprachlichen Korpora; sie müssen also einen hohen „P-Wert“ haben!

Stopwörter hingegen (an denen wir nicht interessiert sind) treten in Fachtexten genauso oft auf wie sonst, werden also bei der Berechnung der Prüfgröße keine besondere Überraschung hervorrufen (d.h. einen niedrigen P-Wert haben). Dasselbe gilt für andere Autosemantika, die z.B. zu anderen Fachsprachen gehören oder nur in der Gemeinsprache auftreten.

Wozu wird also noch die absolute Frequenz des Wortes im Fachtext gebraucht? Hier geht es darum, eine Mindestfrequenz festzulegen, die jedes Wort überschreiten muß, um als Terminus akzeptiert zu werden. Das ist aus mindestens drei Gründen sinnvoll:

- Zum Ausschluß von Rechtschreibfehlern und seltenen Eigennamen, die in der Regel nur einmal vorkommen.
- Für die Repräsentativität: die gefundenen Terme sollen wesentliche Bestandteile der Fachsprache sein.
- Zur Reduzierung des statistischen Fehlers: die relative Häufigkeit des Wortes im Text ist nur eine Schätzung für seine Auftretenswahrscheinlichkeit in Texten der gegebenen Fachdomäne und somit mit einem statistischen Fehler behaftet. Dieser wird kleiner, je größer die absolute Frequenz des Wortes im Text ist. Natürlich sinkt der Fehler auch

mit der Größe der Stichprobe, lange Fachtexte sind also kurzen vorzuziehen.

Ich werde später ein statistisches Verfahren zur Extraktion von Fachtermini entwickeln, das Wörter mit hohem P-Wert und einer gewissen Mindestfrequenz extrahiert (vgl. hierzu [Heyer 2002]). Ein solches Verfahren wird auch Differenzanalyse genannt, da man versucht, die Mengendifferenz des fachbezogenen und des gemeinsprachlichen Wortschatzes zu bilden. Man betrachte auch wieder Abb. 1.1 zur Veranschaulichung: gesucht werden diejenigen lexikalischen Elemente einer Fachsprache, die außerhalb des gemeinsprachlichen Bereichs liegen.

1.4 Linguistische Charakteristika von Fachtermini

Die linguistischen Eigenschaften von Fachtermini sind nur teilweise sprachunabhängig, d.h man muß sie für jede Sprache getrennt herausarbeiten.

Da es in dieser Arbeit um Terminologie im Deutschen und Englischen geht, müssen einige Dinge getrennt voneinander behandelt werden: im Deutschen besteht die Möglichkeit der Komposition zweier Wörter zu einem neuen Wort, während im Englischen zwar auch Wörter zu einem komplexen Term kombiniert werden können, dieser neue Term dann aber fast immer aus mindestens zwei Wörtern besteht.

Die folgende Zweiteilung in Einwort- und Mehrworttermini soll keine Unterscheidung Deutsch-Englisch sein. Aber der Schwerpunkt wird später im Deutschen auf (morphologisch komplexen) Einworttermen, im Englischen auf Mehrworttermen liegen.

1.4.1 Einwortterme

Begriffe der Morphologie

Bevor ich die linguistischen Eigenschaften von Einworttermen untersuche, möchte ich auf einige Grundbegriffe eingehen, die zum weiteren Verständnis benötigt werden:

Gegenstand der Morphologie ist das *Morphem* — das kleinste bedeutungstragende Element der Sprache (vgl. [Bußmann 2002]) — und die Art

und Weise, wie Morpheme zu Wortformen kombiniert werden.

Zur Bildung von Wortformen dienen vor allem die drei Operationen Flexion, Derivation und Komposition. Die letzteren beiden sind interessante Aspekte bei der Extraktion von Fachtermini. Flexion hingegen wird eher als störend empfunden: sie dient der Ableitung grammatischer Vollformen aus einer Grundform und schafft somit syntaktische Oberflächenvarianten ein und desselben Wortstammes. Man neigt dazu, Flexion zu neutralisieren, da man sich eben auf Wortstämme konzentrieren möchte und deren Varianten zu einem Konzept zusammenfassen möchte.

Es gibt mehrere verschiedene Einteilungen von Morphemen. Eine davon berücksichtigt ihre semantische Funktion:

lexikalische oder Basismorpheme sind solche, die Sachverhalte der außersprachlichen Welt bezeichnen. Zu ihnen gehören Nomina wie „Kind“ und „Mantel“, aber auch Verbstämme wie „seh-“

grammatische Morpheme oder Flexive tragen nur grammatische Bedeutung. Sie werden an Wortstämme angefügt (affigiert) und haben z.B. die Form „-t“ in „geh-t“ oder „-er“ in „Kind-er“.

Derivationsaffixe oder Derivative werden ebenfalls an Wortstämme affigiert (man unterscheidet zwischen Primär- und Sekundärstamm, je nachdem, wieviele Affixe dem Stamm schon hinzugefügt wurden). Der Prozeß der Derivation wird unten genauer beschrieben. Derivative sind z.B. „-bar“ in „lös-bar“ oder “un-“ in „un-lösbar“.

Eine weitere Unterscheidung von Morphemen erfolgt danach, ob sie alleine auftreten können oder nicht: *freie* Morpheme wie z.B. „Wort“ oder „rot“ können ohne Affixe im Text vorkommen, *gebundene* Morpheme dagegen nicht. Zu den gebundenen Morphemen zählen alle grammatischen Morpheme und Derivative, aber auch einige Basismorpheme, z.B. „seh-“.

Schließlich gibt es noch den Begriff der Allomorphie: Allomorphe sind die verschiedenen Aussprachevarianten eines Morphems. Beispielsweise taucht der Verbstamm „sprech-“ in den Varianten „sprich-“, „sprach-“ und „spräch“ auf.

Allomorphie stellt ein ernsthaftes Problem dar, wenn man Wörter auf ihre Grundform reduzieren möchte: bei dem Wort „sprichst“ reicht es nicht, das Flexiv „-st“ abzuschneiden, um die Grundform „sprech-“ zu erhalten. Grundformenreduktion ist — wie wir sehen werden — für das Deutsche wesentlich schwieriger als für das Englische, da es im Deutschen wesentlich mehr Flexive, aber eben auch mehr Allomorphe gibt.

Einteilung der Einwortterme

Laut Hoffmann (vgl. [Hoffmann 1988], S. 104) lassen sich Einworttermini einteilen in Simplizia, Derivate, Komposita und Abkürzungen. Ich will deren Eigenschaften im Einzelnen untersuchen:

Simplizia: Hierbei handelt es sich um Wörter, die morphologisch nicht komplex sind und daher auch eine nicht-zusammengesetzte Bedeutung haben. Sie scheinen also zunächst keine hervorstechenden linguistischen Eigenschaften zu haben. Es lassen sich jedoch einige interessante Feststellungen machen:

- *Wortarten:* bei weitem die meisten Einworttermini sind Nomina. Das erscheint logisch, wenn man sich ins Gedächtnis ruft, daß Terme Benennungen für konkrete Begriffe sind (s. Definition Terminologie); diese Begriffe sind meist etwas Gegenständliches, Faßbares und werden somit meist durch Substantive beschrieben.

  Hinzu kommen viele Adjektive mit differenzierender Funktion (vgl. [Fluck 1985], S. 48), d.h. solche, die andere, nominale Terme in ihrer Bedeutung einschränken. Meistens entsteht dabei ein Mehrwortbegriff, der auch zu extrahieren ist (z.B. „*kinetische* Energie“).

  Verben und Adverbien sowie selbstverständlich alles, was wir oben als „Stopwörter“ eingeführt haben, treten nur sehr selten als Einwortterme in Erscheinung.

- Bei *Terminologisierungen* handelt es sich um gemeinsprachliche Wörter, denen eine neue, fachsprachliche Bedeutung zugewiesen

wird. Oft geschieht dies aufgrund von Ähnlichkeit neuer fachbezogener Phänomene mit vorhandenen Dingen aus dem Alltagsleben. Ein Beispiel für Terminologisierung ist der Begriff „Wurzel“, der in der Zahnmedizin und Mathematik neue Bedeutungen erhalten hat.

- *Entlehnungen* spielen vor allem in Gebieten eine Rolle, in denen oft technische Neuerungen aus anderen Ländern importiert werden: deren Name wird einfach aus der fremden Sprache übernommen. Dies läßt sich vor allem in der Computerbranche beobachten: „Server“, „Login“ usw.
- *Lehnübersetzungen* sind hingegen wörtliche Übersetzungen fremdsprachiger Termini, die anfangs oft seltsam anmuten („Ein-Aus-Tastung“), irgendwann aber gar nicht mehr als Lehnübersetzungen wahrgenommen werden: „Luftbild“ von „air photo“. (Zu den Beispielen vgl. [Fluck 1985], S. 54)

Derivate: sind Wortneubildungen, die durch Affigierung von Derivativen an Wortstämme entstehen. Die meisten Derivative haben dabei eine charakteristische Bedeutung, die sie dem Stamm hinzufügen. Derivationssuffixe ändern zudem meist die Wortart des Stamms.

Dabei ist es interessant, daß manche Derivationssuffixe in bestimmten Fachsprachen besonders produktiv sind; im Deutschen ist beispielsweise in vielen technischen Fachsprachen das Suffix „-er“ sehr häufig anzutreffen: „Schweißer“, „Verstärker“, „Absorber“ etc.

Ein weiteres interessantes Phänomen, das der Derivation gleicht, ist die *Konversion*, d.h. die Überführung von Wörtern in eine andere Wortart ohne Anhängen von Derivativen. Ein Beispiel hierfür ist die Nominalisierung „das Schmelzen“; oft werden auch Namen konvertiert und dann als Einheiten benutzt: „Celsius“, „Hertz“ etc.

Komposita: sind ebenfalls Wortneubildungen, bei denen diesmal aber zwei Stämme (bzw. zwei freie Morpheme, vgl. [Bußmann 2002], S. 360) aneinandergefügt werden.

Bei Komposita gibt es stets einen Kopf, der die Wortart des Ganzen bestimmt und dessen Bedeutung (fast immer) durch die der anderen Teile — der sog. Modifikatoren — eingeschränkt wird. Auf diese Weise entstehen Begriffshierarchien, da der Kopf eines Kompositums meist ein Oberbegriff (Hyperonym) des Kompositums ist: ein „Schweröl“ ist ein Öl, eine „Ansauglufttemperatur“ ist eine Temperatur usw.

Statistische Untersuchungen (vgl. [Fluck 1985] S. 49) haben ergeben, daß die meisten Komposita aus nur zwei Teilen bestehen; häufige Muster sind dabei: die Zusammensetzung zweier Nomina („Konjunkturbarometer“), eines Adjektivs mit verschiedenen Köpfen („Leichtöl“, „feingemahlen“) und eines Nomens mit einem Verb („sandstrahlen“, „farbabweisend“). Oft werden auch Abkürzungen miteinbezogen — es entstehen Wörter wie „DLG-prämiert“ oder „EU-Kommission“.

Abkürzungen: Interessant sind hier vor allem Akronyme, die einen Mehrwortbegriff oder ein Kompositum auf einige wenige Buchstaben reduzieren. So wird z.B. aus einer „North Atlantic Treaty Organisation“ einfach die „NATO“ oder aus einem „Personenkraftwagen“ der Pkw. Dies erhöht die Sprachökonomie und garantiert trotzdem weitgehende Eindeutigkeit. Daher sind Abkürzungen in Fachsprachen sehr beliebt, auch wenn sie für Laien oft unverständlich sind.

Ableitung von Heuristiken:

Wie wir gesehen haben, ist ein wichtiges linguistisches Kriterium bei der Suche nach Fachtermini die Wortart der Kandidaten. Mit Hilfe eines sogenannten Part-of-Speech-Taggers, eines Programms, welches jedem Wort eines Textes seine Wortart zuordnet, kann man zunächst die Suche bei Einworttermen auf Nomina und Adjektive einschränken.

Weiter läßt sich die unterschiedliche Produktivität von Derivationssuffixen in den einzelnen Fachsprachen ausnutzen: wenn man weiß, daß manche Suffixe in einer bestimmten Fachsprache statistisch signifikant häufiger auftreten als in der Gemeinsprache (vgl. [Hoffmann 1988] S. 98), so lassen sich diese wieder mittels einer Differenzanalyse herausfiltern. Dann kann man sie dazu benutzen, weitere interessante Einworttermini aufzufinden. Dabei ist

es letztlich auch egal, ob ein Suffix nun wirklich ein Derivativ ist oder nicht. Dies soll an einem Beispiel illustriert werden:

In Naturwissenschaften wie Chemie, Biologie oder Medizin beobachtet man ein gehäuftes Auftreten domänenspezifischer lateinischer oder griechischer Suffixe wie z.B. „-itis“ in der Medizin oder „-ase“ in der Chemie. Diese treten in der Gemeinsprache so gut wie nicht auf und können daher bei einer Differenzanalyse leicht gefunden werden.

Wenn man nun in einem medizinischen Text alle Wörter heraussucht, die auf „-itis“ enden (itis = Entzündung), so werden diese zu 100% Fachtermini sein. Diese Heuristik wird sprachübergreifend funktionieren, da die statistischen Verteilungen der Suffixe in allen Sprachen die oben genannten Auffälligkeiten zeigen.

Nur im Deutschen ausnutzen läßt sich hingegen der hohe Anteil von Komposita unter den Einworttermini. Aufgrund dieses erhöhten Anteils werden Komposita einerseits gute Kandidaten für Terme sein, andererseits lassen sich leicht Termini finden, wenn man eine Liste domänenspezifischer Basismorpheme hat und nach Komposita sucht, die diese enthalten.

Um domänenspezifische Basismorpheme zu finden, kann man z.B. alle Wörter eines Textes einem Kompositazerleger übergeben: dieser identifiziert Komposita und ihre Teile, d.h. diejenigen Basismorpheme, aus denen sich das Kompositum zusammensetzt. Morpheme, die dabei im gesamten Text sehr häufig auftreten, können dann als domänenspezifisch angesehen werden (vgl. hierzu [Heid 1998]). Auch hier ließe sich wieder an eine Differenzanalyse denken.

1.4.2 Mehrwortterme

Zunächst soll geklärt werden, was eigentlich ein Mehrwortterm ist. Es gibt viele aus mehreren Wörtern bestehende Einheiten, deren interne Struktur aber oft unterschiedlich ist.

Zuallererst muß festgestellt werden, daß im Englischen sehr viele aus mehreren Wörtern bestehende Einheiten das bezwecken, was im Deutschen durch Komposita ausgedrückt wird. Wüster (vgl. [Wüster 1991] S. 38) nennt diese Einheiten „Scheingruppen“ und unterscheidet sie von den „echten“

Wortgruppen, deren interne syntaktische Struktur eine andere ist, wie er betont.

Die Schemata, nach denen englische „Scheingruppen“ gebildet werden, entsprechen im Wesentlichen den oben genannten Bildungsmustern für deutsche Komposita, d.h. sehr häufig sind Zusammensetzungen aus zwei Nomina oder aus einem Adjektiv und einem Nomen etc. Einige dieser Scheingruppen sind als Lehnwörter auch in das Deutsche übernommen worden, wie z.B. der Begriff „Information Retrieval“.

Wie lassen sich demgegenüber aber nun die „echten“ Wortgruppen (nun wieder sprachübergreifend) charakterisieren? Betrachtet man diese Frage aus dem Blickwinkel eines Übersetzers, so lautet die Antwort: alles, was als Einheit übersetzt werden muß, ist ein Mehrwortterm (egal, ob Schein- oder echte Wortgruppe). Da ich aber zunächst von einsprachigen Texten ausgehe, ist diese Definition wenig hilfreich für die Extraktion, man sucht also z.B. nach etwas statistisch Verwertbarem.

Wieder lohnt es sich zurückzublicken auf Ferdinand de Saussures Strukturalismus: ein wichtiger von ihm geprägter Begriff ist der des Syntagmas bzw. der syntagmatischen Beziehung (vgl. [Saussure 1967] S. 147 ff.) Syntagmatische Beziehungen sind Kombinierbarkeitsrelationen, d.h. zwei Wörter bilden ein Syntagma, wenn sie nebeneinander im Text auftreten können.

Alle uns interessierenden Mehrworttermini sind also offensichtlich Syntagmen. Weitet man nun den Begriff des Syntagmas aus, indem man fordert, daß die Elemente eines solchen nicht nur gelegentlich, sondern oft (und zwar statistisch signifikant) nebeneinander auftreten, so ergeben sich sogenannte „statistische Syntagmen“ (vgl. [Bordag 2003]).Diejenigen Elemente, die zu einem gegebenen Wort in einer solchen Relation stehen, nennt man auch Nachbarschaftskollokationen des Wortes (vgl. hierzu [Quasthoff 2002]).

Rechnet man statistische Syntagmen aus, so ergeben sich verschiedene Arten von Mehrworteinheiten: es treten *Dependenzen* auf, die auf den semantischen Selektionsbeschränkungen einzelner Wörter beruhen, z.B. „Hund bellt“. Diese sind für uns eher uninteressant, da sie meist keine Fachterminologie darstellen.

Weiter treten *Aufzählungen* als statistische Syntagmen auf, wie z.B. in „Fakultät für Mathematik und Informatik“. Dies können interessante Terme

sein, oft sind aber nur die Teile Fachtermini.

Als nächstes sind *idiomatische Wendungen* zu nennen wie „ins Gras beißen“ oder auch „mit freundlichen Grüßen“, die aber ein gemeinsprachliches Phänomen darstellen: da bei Wendungen wie „ins Gras beißen“ eine einfache Sache (nämlich sterben) mit vielen Wörtern umschrieben wird, ist diese unökonomische Art der Benennung in Fachsprachen sehr selten anzutreffen.

Verbleiben also noch die wirklich interessanten statistischen Syntagmen: Hier sind *Eigennamen* zu nennen, die aus mehreren Teilen bestehen („Deutsche Bank AG“) oder eine Kategorie- bzw. Funktionsangabe gepaart mit einem Eigennamen („Finanzminister Eichel“).

Und schließlich solche Zusammensetzungen aus einem Kopf und einem Modifikator, die nicht als Kompositum realisiert werden (können): hierbei handelt es sich meist um einen nominalen Kopf, der durch ein domänenspezifisches Element — meist ein Adjektiv — in seiner Bedeutung eingeschränkt (modifiziert) wird. Beispiele hierfür sind „kinetische Energie“ oder „gewählter Vertreter“.

Ableitung von Heuristiken

Einerseits stellen wir fest, daß die interessanten Mehrwortbegriffe (egal ob englische Scheingruppen oder echte Wortgruppen in beiden Sprachen) allesamt ganz bestimmte Typen von Nominalphrasen sind.

Das bedeutet, daß sich sogenannte Part-of-Speech-Muster (POS-Muster) bestimmen lassen, denen alle Mehrworttermini entsprechen. In [Arppe 1995] findet man eine quantitative Untersuchung zu diesem Thema, die zu dem Ergebnis führt, daß im Englischen ca. 60% aller Terme eines Textes einem der Muster „Nomen-Nomen“ (N N), „Adjektiv-Nomen“ (A N) und „Nomen“ (N) entsprechen.

Wir können also einen POS-Tagger auf einen Text anwenden und dann die oben genannten Muster mit Hilfe regulärer Ausdrücke herausfiltern.

Dabei werden wir andererseits feststellen, daß wir zwar einen Großteil der relevanten Mehrwortbegriffe finden, aber auch eine Vielzahl von Syntagmen, die nicht statistisch sind (also Wörter, die nur einmal zufällig nebeneinander stehen) und somit keine wirklichen linguistischen Einheiten bilden.

Bourigault drückt es sehr treffend aus:

> „It is possible to devise an extraction program solely based on syntactic data [...] It is not possible to expect this program to extract terminological units *and nothing else*.“ ([Bourigault 1992])

Es erscheint also vielversprechender, die Suche auf diejenigen POS-Muster einzuschränken, welche statistische Syntagmen darstellen, deren Teile also bevorzugt und oft nebeneinander auftreten.

Dieser innere Zusammenhalt der Mehrwortterme wird in der Literatur auch oft „Unithood“ genannt (vgl. z.B. [Kageura 1996]); er ist eine statistische Maßzahl und kann daher leicht berechnet werden und somit eine Art Filter für extrahierte POS-Muster bilden. Auch hier ist der Übergang zwischen statistischen Syntagmen und „zufälligen“ Syntagmen fließend, d.h. es muß ein geeigneter Schwellenwert gefunden werden, um beide voneinander zu trennen.

Weitere Filter sind denkbar, so liegt z.B. die Vermutung nahe, daß Mehrwortterme, die bereits erkannte Einworttermini als Teil enthalten, „besser“ sind, da der Einwortterm schon eine gewisse Fachspezifik mitbringt. Diese Fachspezifik wird oft auch „Termhood“ genannt.

Probleme

Zwei sprachliche Phänomene schränken die Anwendbarkeit der gerade abgeleiteten Heuristik ein:

- Die sogenannte *formale Variation* von Termen ist ein Vorgang, bei dem Phrasen umgestellt werden, so daß die aufgestellten POS-Muster nicht mehr passen. Dadurch kann auch nicht mehr erkannt werden, daß es sich vielleicht um eine bereits extrahierte Phrase handelt (was wiederum die Statistik durcheinanderbringt).

  So ist es z.B. denkbar, daß aus einem „book review“ an anderer Stelle im Text ein „review of books“ wird, wobei sich noch die Frage stellt, ob es sich dann wirklich beide Male um denselben Term handelt. Ich werde zunächst davon ausgehen, daß diese Variation hinreichend selten auftritt, um meine Arbeit empfindlich zu stören.

Bei genormten Terminologien ist Variation natürlich ausgeschlossen, da durch sie Synonyme entstehen, die vermieden werden sollen: Ein Begriff soll nur *eine* Benennung haben.

- Die *Trennbarkeit* mancher Wortgruppen, d.h. die Möglichkeit, die Teile des Mehrwortterms im Satz auseinanderzureißen. Dadurch schlagen die POS-Muster, welche auf den Term passen sollen, natürlich nicht mehr an. Hier läßt sich einwenden, daß die meisten Mehrworttermini Nominalphrasen sind, welche weder im Deutschen noch im Englischen getrennt werden können.

Diese Probleme sollen zunächst nicht beachtet werden. Später wird dann zu prüfen sein, ob dies berechtigt ist, d.h. ob in konkreten Anwendungen die beiden Phänomene wirklich hinreichend selten auftreten, um ignoriert werden zu können.

1.4.3 Semantische Relationen zwischen Termen

Strukturalistische Semantik

Wie wir bereits im letzten Abschnitt gesehen haben, lassen sich viele Ideen des von Saussure begründeten Strukturalismus so präzisieren, daß man sie statistisch gut erfassen kann.

In der Semantik — der Lehre von der *Bedeutung* von Morphemen, Wörtern und Sätzen — gibt es viele verschiedene Richtungen, welche die Bedeutung von Wörtern an verschiedenen Orten suchen: die *kognitive Semantik* vermutet Bedeutungen in unseren Köpfen, die *referentielle Semantik* in der Welt draußen.

Der Vorteil der *strukturalistischen Semantik* liegt darin, daß sie die Bedeutung von Wörtern in der Sprache selbst sucht. Saussure sagt: „Die Sprache ist ein Objekt, das man gesondert erforschen kann“ ([Saussure 1967], S. 17), d.h. man muß nicht unser Gehirn oder die Umwelt erforschen, um Sprache zu verstehen.

Das ist deswegen so vorteilhaft für die Zwecke der Informatik, weil man somit Sprache anhand von Texten alleine untersuchen kann (sofern diese in elektronischer Form vorliegen).

Worin liegt nun also die Bedeutung eines Wortes? Saussure schreibt über das Wort:

> „[...] man muß es auch noch vergleichen [...] mit anderen Wörtern, die man daneben setzen kann; sein Inhalt ist richtig bestimmt nur durch die Mitwirkung dessen, was außerhalb seiner vorhanden ist.“ ([Saussure 1967], S. 138)

Das bedeutet, daß die Bedeutung eines Wortes sich aus seiner Beziehung zu anderen Wörtern ergibt, bzw. aus dem Stellenwert, den es innerhalb des *Systems* der Sprache einnimmt.

Wittgenstein geht in seinen *Philosophical investigations* noch ein Stück weiter: er meint, daß die Bedeutung eines Wortes (zumindest in vielen Fällen) allein durch seinen Gebrauch in der Sprache gegeben ist: „the meaning of a word is its use in the language“ ([Wittgenstein 1978], S. 20, §43).

Aufbauend auf diesen Grundannahmen werde ich nun untersuchen, welche semantischen Relationen zwischen Termen existieren, welche davon sich für die Terminologie-Extraktion ausnutzen lassen und wie man dies realisieren kann.

Paradigmatische Relationen

Neben syntagmatischen Beziehungen erwähnt Saussure noch „assoziative“ Beziehungen zwischen Wörtern. Diese bestehen seiner Meinung nach im Geiste: wir assoziieren bestimmte Wörter miteinander, da sie einander entweder in der Form oder in der Bedeutung ähnlich sind (vgl. [Saussure 1967], S. 147 ff.).

In der modernen Linguistik ist man dazu übergegangen, diese assoziativen Relationen „paradigmatisch“ zu nennen (vgl. [Bußmann 2002], S. 494).

Damit macht man deutlich, daß es sich um Austauschbarkeitsbeziehungen handelt: Wörter, die zueinander in paradigmatischer Beziehung stehen (und somit ein Paradigma bilden), können in bestimmten syntaktischen oder semantischen Umgebungen aufgrund ihrer Ähnlichkeit gegeneinander ausgetauscht werden, genauer gesagt: sie können alternativ an derselben Stelle in einem Syntagma auftreten.

Die meisten interessanten semantischen Beziehungen zwischen Wörtern sind paradigmatischer Natur:

Die offensichtlichste paradigmatische Relation bilden die *Synonyme*: sie sind in beliebigen Kontexten stets gegeneinander austauschbar (man kann in beinahe jedem beliebigen deutschen Satz das Wort *Fahrstuhl* durch das Wort *Lift* oder *Aufzug* ersetzen).

Ein *Hyponym* steht in paradigmatischer Beziehung zu seinem *Hyperonym*: der Begriff *Sperling* läßt sich in den meisten syntaktischen und semantischen Umgebungen sinnvoll durch seinen Oberbegriff *Vogel* ersetzen (dies kann z.B. zur Vermeidung von Wortwiederholungen in Texten dienen).

Umgekehrt werden innerhalb eines Satzes oft Oberbegriffe genannt, die dann durch ihre Hyponyme näher spezifiziert werden: „eine Reihe von *Werkzeugen*, wie z.B. der *Hammer*...“

Auch *Kohyponyme* stehen insofern in paradigmatischer Beziehung zueinander, als sie sich zumindest in beliebigen syntaktischen Umgebungen gegeneinander austauschen lassen; dasselbe gilt für *Antonyme*.

Eine weitere Form von Paradigmen bilden *Wortfelder*: ihre Elemente sind zwar nicht in beliebigen Kontexten gegeneinander austauschbar, dafür besteht aber eine starke assoziative Beziehung zwischen ihnen (z.B. Wortfeld „Wirtschaft“: *Arbeit, Unternehmen, Aktien, Geld,...*).

Schließlich bezeichnet man noch — aufgrund ihrer syntaktischen Austauschbarkeit — alle Wörter derselben *Wortart* (oder auch Flexionsklasse) als Paradigma.

Semantisch gesehen ist das Wortfeld die weiteste Form des Paradigmas: es umfaßt Wörter aus einem bestimmten Bedeutungsbereich, die „lückenlos einen bestimmten begrifflichen oder sachlichen Bereich abdecken sollen“ ([Bußmann 2002], S. 753) und die zueinander wieder in Relationen wie Synonymie, Antonymie, Hyponymie oder Kohyponymie stehen.

Da Terminologie, wie oben gesehen, alle Fachausdrücke eines Fachgebietes umfaßt, die allgemein üblich sind, kann man sagen: Die Terminologie

einer Fachdomäne ist nichts anderes als ein Wortfeld, das den sachlichen Bereich eben dieses Fachgebiets abdeckt.

Präzisierung für die Terminologie-Extraktion

Haben wir nun einen Teil der Terminologie aus einem Fachtext extrahiert, so müßte sich diese erweitern lassen, indem wir alle weiteren Begriffe auffinden, die zusammen mit den bereits gefundenen ein Wortfeld bilden, also irgendwie zu diesen in paradigmatischer Beziehung stehen. Die Frage ist, ob sich Wortfelder irgendwie berechnen lassen.

Ein erster Ansatz könnte sein, das gemeinsame Vorkommen der Elemente eines Wortfeldes in bestimmten Umgebungen auszunutzen: oft tauchen sie in Aufzählungen auf („Hammer und Meißel“) oder es werden — wie oben erwähnt — Oberbegriffe durch Nennung einiger Hyponyme spezifiziert.

Darüber hinaus wäre es aber hilfreich, den Begriff der paradigmatischen Beziehung so zu präzisieren wie oben den des Syntagmas, d.h. eine Art „statistische paradigmatische Beziehung“ zu definieren.

Nimmt man — wie Wittgenstein — an, daß die Bedeutung (oder zumindest eine gute Charakterisierung der Bedeutung) eines Wortes in dessen Gebrauch in der Sprache liegt, so läßt sich sagen: zwei Wörter sind dann semantisch ähnlich bzw. gehören demselben Wortfeld an bzw. stehen in paradigmatischer Beziehung, wenn sie in ähnlichen Kontexten gebraucht werden.

Der Kontext eines Wortes A läßt sich aber nun wieder mit statistischen Mitteln erfassen: beispielsweise durch die Menge seiner Satzkollokationen, d.h. durch die Menge derjenigen Wörter, die häufig (und zwar statistisch signifikant) zusammen mit A in ein und demselben Satz auftreten (vgl. hierzu [Heyer 2001], Abschn. 4.3 und [Biemann 2003]).

In [Bordag 2003] findet man daher folgende Definition der paradigmatischen Beziehung:

> „Die paradigmatische Relation [...] zwischen zwei Wortformen ist genau dann gegeben, wenn die globalen Kontexte der Wortformen ähnlich einander sind.“

Unter dem globalen Kontext einer Wortform wird dabei eben die Menge ihrer Satzkollokationen verstanden.

Ableitung einer Heuristik:

Zunächst liegt es nahe, wieder über entsprechende POS-Muster in Fachtexten nach Aufzählungen zu suchen, welche bereits gefundene Terme enthalten. Die weiteren in der Aufzählung vorkommenden Wörter sind dann vermutlich auch Fachtermini.

Desweiteren haben wir gesehen, daß sich der Gebrauch (und damit die Bedeutung) eines Terms A durch seine Satzkollokationen charakterisieren läßt und daß sich somit ähnliche Wörter zu A finden lassen, indem man nach Elementen mit ähnlichen Satzkollokationen sucht.

Man könnte also eine, z.B. durch Differenzanalyse gefundene Menge von Fachtermini erweitern, indem man nach weiteren Termen sucht, die zu diesen in „statistischer“ paradigmatischer Beziehung stehen.

Dazu müßte man zu jedem Wort eines Fachtextes die stärksten Satzkollokationen in Form eines Vektors bestimmen und dann ein geeignetes Ähnlichkeitsmaß zwischen Vektoren festlegen. Daraus ergibt sich ein Algorithmus, wie er in Tabelle 1.1 beschrieben ist.

```
1 Bestimme zu jedem Wort w die Menge seiner
  Satzkollokationen in Form eines Vektors S(w)
2 Bestimme eine erste Auswahl von Fachtermini F
3 Zu jedem Term x aus F
   Bestimme alle Terme b mit: S(b) ähnlich S(x)
   Falls b nicht in F: erweitere F um b
4 Falls noch Terme zu F hinzugekommen sind, gehe zu 3
```

Tabelle 1.1: Algorithmus zur Extraktion ähnlicher Begriffe zu einer gegebenen Ausgangsmenge von Termini

Mit diesem Verfahren müßte sich der *Recall* (die Ausbeute) der Termi-

nologie-Extraktion deutlich verbessern lassen.

Die schrittweise Erweiterung der Menge **F** läßt sich dabei auch als eine Form des Relevance Feedback[1] realisieren, falls nämlich der Anwender nach jedem Iterationsschritt die für ihn relevanten Terme auswählt.

Abzuwarten bleibt noch, wie sich die *Precision* (die Genauigkeit) bei diesem Verfahren entwickelt. Wichtig wird in diesem Zusammenhang sein, wie man den Schwellenwert festlegt, ab dem zwei Vektoren aus Satzkollokationen als ähnlich gelten.

1.5 Ideen aus dem Information Retrieval

Die Forschung im Gebiet Information Retrieval beschäftigt sich mit dem Auffinden von Information in unstrukturierten, natürlichsprachlichen Dokumenten.

In einem typischen Anwendungsfall (als *Suchmaschine* bekannt) gibt ein Anwender einen oder mehrere Suchterme ein und erhält als Antwort eine Liste von Dokumenten, welche die Suchterme enthalten. Diese sollen dabei meist noch nach Relevanz geordnet sein, d.h. dasjenige Dokument, das bezüglich der Anfrage am wahrscheinlichsten relevant ist, soll in der Liste ganz oben stehen.

Eine wichtige Vorbereitung für das Suchen ist die Auswahl von Schlagwörtern zur Beschreibung von Dokumenten und die Berechnung von Gewichten für diese Schlagwörter. Dieser Vorgang wird *automatisches Indexieren* genannt und soll im Folgenden beschrieben werden.

1.5.1 Automatisches Indexieren

Das automatische Indexieren ist eine Aufgabe, die der Extraktion von Terminologie in gewisser Hinsicht ähnlich ist. Es handelt sich dabei um die Repräsentierung von Dokumenten durch eine Menge von Indextermen (oft spricht man von Deskriptoren). Die Indexterme sollen dabei den Inhalt des jeweiligen Dokumentes möglichst gut wiedergeben und helfen, das Dokument von anderen zu unterscheiden.

[1] siehe Abschnitt 1.5.2

Die Frage lautet also hier: „Welche Terme sind besonders charakteristisch für das vorliegende Dokument?“ Handelt es sich bei dem Dokument um einen Fachtext, so wird aus dieser Problemstellung die Frage nach Termen, die besonders charakteristisch für die Fachdomäne des Textes sind, d.h. die Indexterme entsprechen ziemlich genau der Fachterminologie.

Der Unterschied zwischen Fragestellungen im Information Retrieval (IR) und den hier behandelten Problemen ist die Tatsache, daß in IR-Anwendungen stets eine Dokumentenkollektion zugrundeliegt, die für die Berechnung von Termgewichten als Vergleichskorpus herangezogen wird.

Die Kollektion ist dabei oft sehr heterogen, d.h. normalerweise kann man nicht davon ausgehen, daß eine ausgewogene Mischung allgemeinsprachlicher Texte vorliegt, mit der ein Fachtext verglichen werden kann.

Es wird also im Allgemeinen nicht die Fachspezifik eines möglichen Indexterms bewertet, sondern seine Seltenheit innerhalb der Kollektion. Diese ist aber eng mit der Fachspezifik verknüpft, denn je seltener ein Term in einer Kollektion ist, desto wahrscheinlicher gehört er einer bestimmten Fachdomäne an, die nur durch einige wenige Dokumente repräsentiert wird. Welche Domäne innerhalb einer Kollektion aber wie stark vertreten ist, ist meist nicht bekannt.

Wie wir in Kapitel 2 sehen werden, sind viele der dort beschriebenen statistischen Ansätze für Aufgabenstellungen im IR-Bereich entwickelt worden. Es lohnt sich also, die Problemstellung des automatischen Indexierens genauer zu studieren.

1.5.2 Relevance Feedback

Eine weitere Technik aus dem IR-Bereich — das sogenannte Relevance Feedback — beschäftigt sich mit dem in Abschnitt 1.2 angesprochenen Problem, die Ergebnisse eines Systems an die Bedürfnisse des Benutzers anzupassen.

Im Information Retrieval formuliert der Benutzer einer Suchmaschine eine Anfrage, die meistens aus einem oder mehreren Suchtermen besteht. Diese sind aber oft nur eine von vielen Möglichkeiten, das zu beschreiben, wonach er sucht, d.h. es ist wahrscheinlich, daß Dokumente existieren, welche die Suchterme des Benutzers zwar nicht enthalten, aber trotzdem relevant

sind, da sie genau das Thema behandeln, das den Anwender interessiert — nur eben mit einer anderen Wortwahl.

Hierzu ein Beispiel: der Benutzer möchte sich über Autos informieren und gibt dementsprechend den Suchbegriff „Auto“ ein; ziemlich sicher interessieren ihn aber auch Dokumente in denen es um „Pkws“ oder „Automobile“ geht. Diese werden in klassischen IR-Systemen nicht gefunden.

Vielen Menschen fällt es schwer abzuschätzen, wie sie ihre Anfrage verfeinern müssen, um mehr und passendere Dokumente zu finden; es ist also wünschenswert, ein Verfahren zu finden, welches den Benutzer bei dieser Arbeit unterstützt.

Der Ansatz des Relevance Feedback (wie er z.B. in [Salton 1990] beschrieben wird) basiert auf der Idee, nach einem ersten, „klassischen“ Suchdurchlauf den Benutzer einige relevante und nichtrelevante Dokumente auswählen zu lassen und seine Anfrage dann wie folgt zu modifizieren: Terme, die in Dokumenten vorkommen, die der Benutzer für relevant hielt, erhalten ein größeres Gewicht, Terme aus nichtrelevanten Dokumenten hingegen ein kleineres. Man benutzt also die Eigenschaften derjenigen Dokumente, die der Benutzer relevant findet, um weitere zu finden.

Dieses Verfahren bietet mehrere Vorteile (vgl. hierzu auch [Salton 1990]):

- Es erhöht den Recall, d.h. die Anzahl der gefundenen relevanten Dokumente, um ein Beträchtliches, da auch Dokumente gefunden werden, welche die ursprünglichen Suchterme nicht enthalten (s.o. „Auto“ und „Pkw“).

- Der Benutzer muß keinen genauen Einblick in die Implementierungsdetails des Systems haben, um seine Anfrage zu verfeinern.

- Der Prozeß des Suchens wird in mehrere Schritte zerlegt, in denen man sich dem Ziel immer mehr annähert, theoretisch — wenn der Benutzer die Geduld hat — bis hin zum erschöpfenden Recall.

Ableitung eines Verfahrens

Wie lassen sich diese Erkenntnisse auf Terminologie-Extraktion anwenden? Wie wir in Abschnitt 1.2 festgestellt haben, hängt es oft von der konkreten

Anwendung, d.h. auch vom Anwender ab, was als Terminologie angesehen wird und was nicht. Es ist also erwünscht, die Ergebnisse der Extraktion an die Nutzerbedürfnisse anzupassen.

Rufen wir uns die prinzipielle Idee des Relevance Feedback ins Gedächtnis: die Eigenschaften bereits gefundener, vom Benutzer als relevant eingestufter Elemente auszunutzen, um weitere gute Elemente zu finden. Dies läßt sich direkt auf die Suche nach Fachtermini übertragen: Dem Anwender werden zunächst einige — z.B. durch eine Differenzanalyse gefundene — Termkandidaten präsentiert und er wählt daraus diejenigen aus, die er für sinnvolle Terminologie hält. Nun gilt es, die Eigenschaften der vom Nutzer ausgewählten Terme zu analysieren und weitere Terme mit ähnlichen Eigenschaften zu finden.

Im Information Retrieval sind die Eigenschaften der gesuchten Elemente Indexterme (Deskriptoren), die Anfrage des Benutzers wird also um gute Deskriptoren erweitert. Es wird später noch zu klären sein, mit welchen Attributen man Fachterme beschreiben sollte, so daß zu einer gegebenen Menge vorklassifizierter Terme die Suche nach Termen mit ähnlichen Attributwerten auch wirklich weitere gute Fachtermini liefert.

Kapitel 2

Bestehende Ansätze

Bevor ich zur Beschreibung eines von mir entworfenen und implementierten Verfahrens komme, möchte ich zunächst einige schon bestehende Ansätze zur Terminologie-Extraktion beschreiben, um einen Einblick in den Stand der Forschung zu geben und darzulegen, welche dieser Ideen mich bei meinem Entwurf beeinflußt haben bzw. welche mir für meine Zwecke unbrauchbar erschienen.

Dabei soll zunächst auf Ansätze aus dem Information Retrieval eingegangen werden: die meisten hiervon sind statistische Verfahren, die das automatische Indexieren als Zielstellung haben. Es sind aber auch ein paar wenige linguistische Ansätze darunter.

Anschließend möchte ich eine Auswahl von Systemen vorstellen, die mit dem ausdrücklichen Ziel der Terminologie-Extraktion entworfen wurden; diese sind meist entweder rein linguistisch motiviert oder verfolgen einen hybriden Ansatz: Wortstatistiken gepaart mit linguistischen Informationen. Innerhalb der einzelnen Abschnitte werden die Verfahren dabei grob chronologisch geordnet.

Insgesamt werden nur Verfahren Erwähnung finden, die entweder grundlegend bzw. richtungsweisend für die Forschung im Bereich des automatischen Indexierens bzw. der Terminologie-Extraktion waren oder die Ideen enthalten, welche ich für meine Zwecke als sehr interessant empfand. Dabei kann die Auswahl natürlich weder vollständig noch vollkommen objektiv sein, so daß man die folgende Abhandlung nur als grobe Orientierung anse-

hen sollte.

2.1 IR-Ansätze zum Automatischen Indexieren

2.1.1 Statistische Verfahren

[Bookstein 1974]

Ein früher Ansatz zum automatischen Indexieren stammt von Bookstein und Swanson (vgl. [Bookstein 1974]). Er basiert auf der Idee, daß signifikante Terme clustern, d.h. daß sie gehäuft in wenigen Dokumenten auftreten:

In einer Dokumentenkollektion wissenschaftlicher Arbeiten zu verschiedenen Themen wird z.B. das Wort *laser* in manchen Dokumenten recht häufig, in anderen dagegen gar nicht auftreten. Das Wort *obtain* hingegen verteilt sich vermutlich relativ gleichmäßig über die Kollektion.

Laut Bookstein korreliert die ungleichmäßige Verteilung des Wortes *laser* direkt mit seiner Brauchbarkeit als Indexterm: es ist charakteristisch und wichtig für den Inhalt derjenigen Dokumente, in denen es vorkommt, während *obtain* keine wesentliche Information über den Inhalt eines Artikels in sich trägt, da es in allen Artikeln mit gleicher Wahrscheinlichkeit vorkommt.

Bookstein führt also ein Maß für die Anhäufung eines Terms in bestimmten Dokumenten ein, den sogenannten Clusterwert x: für ein gegebenes Wort w ist dieser gegeben durch

$$x(w) = \sum_j \binom{j}{2} s_j(w) = s_2 + 3s_3 + 6s_4 + ... \tag{2.1}$$

wobei s_j die Anzahl der Dokumente angibt, in denen w genau j-mal auftritt. Bookstein wählt dieses Maß in Analogie zu chemischen Bindungen: tritt ein Wort mehrfach innerhalb eines Artikels auf, so besteht zwischen den einzelnen Vorkommen jeweils eine Bindung, die belohnt werden soll. Bei j Vorkommen innerhalb des Dokuments macht das $\binom{j}{2}$ Bindungen. Das Maß x wird also groß, wenn ein Wort in wenigen Artikeln auftritt, dort aber jeweils häufig.

Weiter geht Bookstein von einer Poisson-Verteilung der Wörter über die Dokumente aus: tritt ein Wort w in A Dokumenten R-mal auf, so ist die

Wahrscheinlichkeit, daß ein beliebiges dieser Dokumente das Wort w genau k-mal enthält:

$$P(k) = \frac{1}{k!}(R/A)^k e^{-\frac{R}{A}} \tag{2.2}$$

mit einem Erwartungswert von R/A. Schließlich berechnet Bookstein für jedes Wort w die Funktion $g(x)$, welche zum Clusterwert $x(w)$ des Wortes w ausgibt, wie wahrscheinlich es unter der Annahme der Poisson-Verteilung des Wortes ist, daß dieser Wert x erreicht oder überschritten wird. $g(x)$ ist also klein, wenn ein Wort stark von der Poisson-Verteilung abweicht, d.h. an wenigen Stellen gehäuft auftritt. Wörter mit g-Werten unterhalb eines bestimmten Schwellenwertes werden demzufolge als gute Indexterme angesehen.

[Salton 1975]

Einen recht ähnlichen Ansatz verfolgt [Salton 1975]: auch er möchte Wörter als Indexterme bevorzugen, die nur in wenigen Dokumenten vorkommen, dort aber häufig sind.

Er wählt dazu die Kombination zweier Maße, die wesentlich einfacher zu verstehen und zu berechnen sind als die von Bookstein verwendeten, und die daher für die Gewichtung von Termen in IR-Systemen große Verbreitung gefunden haben: TF (term frequency) und IDF (inverse document frequency). Die TF f_{ik} mißt dabei, wie häufig Term k im Dokument i vorkommt, während die IDF_k eines Terms k berechnet wird als:

$$IDF_k = log\frac{N}{d_k} + 1 \tag{2.3}$$

N ist dabei die Gesamtzahl der Dokumente in der Kollektion, d_k die Anzahl der Dokumente, in denen Term k auftritt. Die IDF wird also groß, wenn Term k in möglichst wenigen Dokumenten auftritt.

Salton verwendet dann das Produkt $w_{ij} = f_{ik} * IDF_k$ als Gewicht für Term k im Dokument i. Das Gewicht w_{ij} ist also sehr ähnlich dem von Bookstein verwendeten Clusterwert. Salton benützt es aber direkt zum Gewichten und verzichtet im Unterschied zu Bookstein auf die aufwendige Berechnung der Wahrscheinlichkeit dieser Werte unter der Annahme einer Poisson-Verteilung.

Terme als interessant auszuwählen, die im vorliegenden Text häufig, sonst aber selten auftreten, entspricht der Idee der in Kapitel 1 beschriebenen Differenzanalyse: die dort erwähnte Mindestfrequenz eines Terms gleicht der TF, da sie versucht, die Repräsentativität des Terms für das vorliegende Dokument zu sichern; der Wert einer statistischen Prüfgröße P eines Wortes im Bezug auf einen Vergleichskorpus entspricht grob der IDF, da Terme bevorzugt werden, die im Referenzkorpus selten sind — so wie bei der IDF Terme, die nur in wenigen Dokumenten auftreten.

Da die Ansätze von Bookstein und Salton aber beide mit Dokumentenkollektionen und nicht mit einem einheitlichen Referenzkorpus arbeiten, lassen sich ihre Formeln nicht direkt auf die Terminologie-Extraktion mit einem Vergleichskorpus anwenden. Sie müssen vielmehr angepaßt werden.

[Salton 1989]

Ein weiterer interessanter Aspekt wird in [Salton 1989] (S. 294 ff.) näher beschrieben: die Verwendung von Phrasen als Indextermen. Die Phrasen werden dabei nach rein statistischen Kriterien gebildet und dienen dazu, die Diskriminanzwerte von Einworttermen zu verbessern: hochfrequente Terme, die in fast jedem Dokument auftreten und somit wenig dazu beitragen, Dokumente voneinander zu unterscheiden (schlecht diskriminieren), möchte Salton um mittel- oder niederfrequente Modifikatoren erweitern. So will er ihre Frequenz senken und ihren Diskriminanzwert steigern. Als Modifikatoren eines gegebenen Kopfes K sollten dabei solche Terme gewählt werden, die häufig mit K zusammen auftreten (also in statistischer syntagmatischer Relation zu K stehen). Ob der dabei entstehende Mehrwortterm tatsächlich eine Phrase im linguistischen Sinne ist, interessiert Salton nur nachrangig.

Die dahintersteckende Idee scheint mir jedoch recht interessant, z.B. als ein Filter für Phrasen, die mittels POS-Mustern extrahiert werden, wie in Abschnitt 1.4.2 beschrieben: Da die Bedeutung eines Kopfes bei der Bildung von Komposita durch die Bedeutung der Modifikatoren eingeschränkt bzw. „verfeinert“ wird, erscheint es logisch, daß im allgemeinen der Kopf des Kompositums eine höhere Frequenz im Text hat als die Modifikatoren. Eine hohe Frequenz des Kopfes spricht auch dafür, daß dieser ein wichtiges Konzept

innerhalb der betrachteten Domäne beschreibt.

[Damerau 1993]

Auch [Damerau 1993] geht davon aus, daß signifikante Terme clustern. Er konzentriert sich auf die Extraktion domänenspezifischer Wort*paare* aus Texten. Grundlage seiner Experimente ist dabei ein Korpus, welches sich aus Texten vieler verschiedener Fachbereiche zusammensetzt.

Zur Extraktion von Zweiwort-Termini einer bestimmten Domäne betrachtet er dann jeweils nur die Texte zu diesem Thema und vergleicht dort gefundene Ergebnisse mit solchen, deren Berechnungsgrundlage das Gesamtkorpus war.

Er verwendet dazu die *association ratio* zweier Wörter x und y:

$$ar(x,y) = \log \frac{P(x,y)}{P(x)P(y)} \tag{2.4}$$

Dieses Maß ist unter dem Namen *Mutual information* bekannter und dient dazu, die statistische Abhängigkeit der beiden Wörter zu messen; es handelt sich also um ein Kollokationsmaß. Die Auftretenswahrscheinlichkeiten $P(x), P(y), P(x,y)$ werden dabei aus dem Korpus geschätzt als relative Häufigkeiten der Wörter x und y bzw. des Wortpaars $< x, y >$.

Mittels der association ratio erhält Damerau also Wörter, die (signifikant) häufig als Paar auftreten. Um nun auszunutzen, daß Fachtermini gehäuft in Texten ihrer Domäne auftreten, sonst aber selten, berechnet er zwei verschiedene association ratios: einmal unter Verwendung der Auftretenswahrscheinlichkeit eines Wortpaars nur in Texten der Fachdomäne ($P_f(x,y)$), einmal unter Berücksichtigung des Auftretens im gesamten Korpus ($P_g(x,y)$). Nun bildet er die Differenz beider association ratios:

$$\begin{aligned} \log \frac{P_f(x,y)}{P(x)P(y)} - \log \frac{P_g(x,y)}{P(x)P(y)} &= \log \left(\frac{P_f(x,y)}{P(x)P(y)} \frac{P(x)P(y)}{P_g(x,y)} \right) \\ &= \log \left(\frac{P_f(x,y)}{P_g(x,y)} \right) \end{aligned}$$

Was nach den Umformungen als Argument des Logarithmus übrigbleibt, ist der Quotient der relativen Häufigkeit des Wortpaars in den Fachtexten geteilt durch seine relative Häufigkeit im Gesamtkorpus. Damerau nennt dieses

Verhältnis *relative frequency ratio*, ich werde es später *Häufigkeitsquotient* (HQ) nennen.

Dieses Maß gibt nun keine Auskunft mehr über die Kollokationsstärke zwischen beiden Wörtern x und y, sondern nur noch über die Fachspezifik der gesamten Phrase. Damerau schreibt aber, er habe damit gute Ergebnisse erzielt. Der Häufigkeitsquotient läßt sich natürlich auch für einzelne Wörter berechnen (im Sinne einer Differenzanalyse), was ich auch getan habe (siehe Abschnitt 3.2).

Das Verfahren verzichtet auf eine statistische Modellierung der Wortfrequenzen (z.B. mittels einer Poissonverteilung wie bei Bookstein). Damerau begründet dies damit, daß solche Modelle annehmen, ein beliebiges Wort sei an jeder Stelle eines Korpus gleichwahrscheinlich. Dies ist aber nicht der Fall, da signifikante Terme clustern, also bevorzugt an bestimmten Stellen eines Textes auftreten.

Eine vertiefende Diskussion zur Frage der statistischen Modellierung einer Differenzanalyse findet sich in Abschnitt 3.2.

[Cohen 1995]

Eine ganz andere Herangehensweise verfolgt [Cohen 1995]: Er wählt Indexterme nach den in ihnen enthaltenen Teilzeichenketten (n-grammen) aus, die charakteristisch für die jeweilige Domäne sind. Dabei ist es ihm wichtig, die Funktionsfähigkeit seines Programmes völlig unabhängig von Sprache und Fachdomäne zu garantieren.

Dieser Ansatz ist sehr ähnlich der in Abschnitt 1.4.1 abgeleiteten Suffix-Heuristik: dort wurde eine Differenzanalyse für Teilzeichenketten nur am Ende von Wörtern vorgeschlagen, Cohen betrachtet Substrings an beliebiger Position im Wort — sein Ansatz ist also etwas allgemeiner.

Um die Sprachunabhängigkeit zu garantieren, verzichtet Cohen auf die Verwendung einer Grundformenreduktion, einer Stopwortliste oder einer syntaktischen Analyse, welche alle notwendigerweise von Sprache zu Sprache angepaßt werden müßten. Zur Extraktion von Indextermen für ein gegebenes Dokument D geht er also wie folgt vor:

- Man repräsentiere D durch einen Vektor, in dem man zu jedem n-

gramm dessen Frequenz im Text speichert.

- Ebenso lege man einen Vektor für einen Vergleichskorpus K an, der zu jedem n-gramm, welches in K auftritt, dessen Frequenz enthält.

- Nun erhält jedes n-gramm g aus D ein Gewicht, welches um so höher ist, je stärker seine Frequenz in D von seiner Frequenz im Vergleichskorpus K abweicht.

- Schließlich werden Wörter ausgewählt, welche viele gute n-gramme (also solche mit hohem Gewicht) in ihrer Mitte enthalten. Der genaue Auswahlprozeß für Wörter ist etwas komplizierter, darauf soll aber hier nicht eingegangen werden.

  Interessant ist aber zu erwähnen, daß n-gramme auch über Wortgrenzen hinaus ausgewertet werden, d.h. daß auch eine Gruppe aus zwei Wörtern ausgewählt werden kann, falls ein oder mehrere „gute“ n-gramme existieren, welche jeweils mindestens einen Buchstaben aus beiden Wörtern (nämlich vor und nach dem Leerzeichen zwischen beiden Wörtern) enthalten und somit beide Wörter umspannen. Auf diese Weise können auch Phrasen extrahiert werden.

Zur Berechnung des Gewichts eines n-grams benutzt Cohen einen statistischen Test, den G^2-Test. Dieser mißt, wie unwahrscheinlich es ist, ein n-gramm g_i im Text C_i-mal zu beobachten unter der Voraussetzung (Nullhypothese), daß die Auftretenswahrscheinlichkeit p_i des n-grams im Text gleich der im Korpus q_i ist. Dabei schätzt er die Wahrscheinlichkeiten p_i und q_i mittels ihrer relativen Frequenz in Text bzw. Korpus ab (*maximum likelihood estimate*):

$$p_i = \frac{C_i}{S}, \quad q_i = \frac{B_i}{R}$$

C_i und B_i sind dabei die absoluten Frequenzen des n-grams in D bzw. K, während S und R die absolute Anzahl der n-gramme in D bzw. K bezeichnet.

Seine Prüfgröße ψ_i sagt nun aus, wie signifikant die Abweichung der gemessenen Frequenz ist, wenn man die Nullhypothese voraussetzt. Da er dabei nur eine Abweichung nach oben (relative Häufigkeit in D größer als

in K, d.h. $p_i > q_i$) berücksichtigen will, gelangt er schließlich zu folgender Formel:

$$\psi_i = \begin{cases} Sp_i \log p_i + Rq_i \log q_i - (S+R)t_i \log t_i & \text{falls } p_i > q_i \\ 0 & \text{sonst} \end{cases} \tag{2.5}$$

mit $t_i = \frac{C_i + B_i}{S+R}$. Zur Extraktion der Indexterme braucht Cohen schließlich noch einen geeigneten Schwellenwert: Wörter erhalten als Gewicht die Summe der Gewichte (ψ_i) aller in ihnen vorkommenden n-gramme und werden ausgewählt, wenn ihr Gewicht einen Schwellenwert übersteigt. Da dieser sich dynamisch an die Textgröße anpassen soll, wählt Cohen ihn als den Mittelwert aller Wortgewichte plus zweimal deren Standardabweichung.

Cohens Verfahren ist nicht nur wegen seiner Sprachunabhängigkeit interessant: die Annahme, daß in manchen Fachdomänen bestimmte Derivationssuffixe besonders häufig sind, läßt sich offensichtlich (wenn man die Ergebnisse Cohens betrachtet) auf n-gramme an beliebigen Positionen im Wort übertragen und somit verallgemeinern. Außerdem schlägt Cohen eine Möglichkeit vor, den Schwellenwert für eine Differenzanalyse dynamisch anzupassen.

Cohens Ansatz hat aber auch einige Nachteile: aufgrund der beschränkten Länge seiner n-gramme (meist $n = 5$) können seine Wortgruppen nur höchstens zwei Wörter umfassen.

Zudem findet sich in seinen Ergebnissen relativ viel Rauschen. Es werden z.B. auch Wortgruppen wie „worden sind“ extrahiert und es finden sich unter seinen Indextermen oft Vollformen derselben Grundform, z.B. „earthquake“ und „earthquakes“, da er ja auf Grundformenreduktion verzichtet.

All diese Probleme lassen sich natürlich recht leicht mittels linguistischer Methoden beheben, was aber wieder auf Kosten der Sprachunabhängigkeit ginge.

2.1.2 Linguistische Verfahren

Auch unter den IR-Ansätzen sind einige, die sich linguistischer Methoden bedienen. Jedes dieser Systeme basiert in irgendeiner Weise auf der Extrak-

tion von Nominalphrasen aus POS-getaggtem oder vollständig geparstem Text.

[Dillon 1983]

[Dillon 1983] beschreibt sein Indexiersystem FASIT (Fully Automatic Syntactically based Indexing of Text), welches auf der Extraktion von POS-Mustern aus einem getaggten Text aufbaut.

In einer ersten Phase werden dabei die Muster (Dillon nennt sie Konzepte) extrahiert, um danach in einem zweiten Schritt in eine sogenannte kanonische Form gebracht und nach semantischer Ähnlichkeit gruppiert zu werden. Die kanonische Form (KF) einer Phrase wird gebildet, indem:

- Präpositionen und Konjunktionen herausgelöscht,
- die übrigbleibenden Phrasenteile auf ihre Grundform reduziert, und schließlich
- diese Grundformen alphabetisch sortiert werden.

Somit wird aus der Phrase „review of books“ die KF „book review“, genauso wie aus „book review“, das sich bei der Transformation in die KF nicht verändert. Dillon behebt somit das in Abschnitt 1.4.2 angesprochene Problem der formalen Variation von Phrasen, und indexiert somit synonyme Phrasen wie die gerade genannten nur einmal.

Um auch erschöpfend zu indexieren und semantisch ähnliche Phrasen zu einer Anfrage zu finden, teilt Dillon die Phrasen zusätzlich (gemäß der Einwort-Stämme, die sie enthalten) in Gruppen ein und verwendet auch die Teile von Phrasen als Indexterme. Interessant ist aber vor allem, mit welch einfachen und effizienten Mitteln er das Problem der formalen Variation löst.

[Salton 1988]

[Salton 1988] beschäftigt sich mit der Erstellung von Indizes für Bücher. Im Gegensatz zu Anwendungen im IR muß man hier — meint Salton — auf jeden Fall Phrasen als Indexterme berücksichtigen. Um passende Nominalphrasen (NPs) als Indexterme zu identifizieren, schlägt Salton eine vollständige syntaktische Analyse des Textes vor. Aus deren Ausgabe möchte

er dann nominale Köpfe extrahieren, zusammen mit den Köpfen aller ihrer Modifikatoren. Die so erhaltenen Phrasen müssen noch gefiltert werden durch:

Einfache Regeln: Längere Phrasen (z.B. aus drei oder mehr Wörtern) sind kürzeren vorzuziehen, Nomen-Nomen-Muster sind besser als Adjektiv-Nomen etc.

TF/IDF: Soll hier berechnet werden, indem man das Buch in Kapitel (oder Abschnitte) unterteilt und diese als Dokumentenkollektion betrachtet.

Die Anwendung des TF/IDF-Maßes auf Kapitel von Büchern geht wieder von der Idee aus, daß signifikante Terme clustern, d.h. gehäuft an einer Stelle des Buches auftreten. Diese Idee ist potentiell auch für Terminologie-Extraktion interessant, scheint aber nur bei sehr großen Texten sinnvoll anwendbar, während das von mir zu entwickelnde Verfahren insbesondere für kleine Texte gute Ergebnisse liefern soll.

[Evans 1995]

Schließlich soll noch der Ansatz von [Evans 1995] Erwähnung finden: sein System CLARIT führt keine vollständige syntaktische Analyse von Texten durch, sondern beschränkt sich auf das oberflächliche Parsen von NPs (auch als Chunking bekannt).

Dabei sollen keine maximalen Nominalphrasen (die auch nachgestellte Modifikatoren wie z.B. Präpositionalphrasen enthalten können), extrahiert werden, sondern nur sogenannte „simplex NPs“, die nur aus einem nominalen Kopf und *vorangestellten* Modifikatoren bestehen.

Ein Beispiel für eine maximale NP ist die Phrase „high density disk drive in modern computers“. Diese wird durch CLARIT in zwei simplex NPs „high density disk drive“ und „modern computers“ zerlegt, die dann als Indexterme verwendet werden.

Der Vorteil eines NP-Chunkings bzw. einer vollständigen syntaktischen Analyse liegt — gegenüber einer einfachen Extraktion von POS-Mustern —

darin, daß keine „zufällig benachbarten“ Elemente mit passenden POS-Tags extrahiert werden, die aber unterschiedlichen Phrasen angehören.

In dem Satz „the man gave the girl biscuits“ entspricht die Wortfolge „girl biscuits“ dem POS-Muster N N (Nomen Nomen) und würde somit als interessante Phrase identifiziert werden. Sie ist aber keine zusammenhängende Nominalphrase: „girl“ und „biscuits“ sind jeweils eigene Phrasen, die als getrennte Konstituenten (indirektes und direktes Objekt) im Satz auftreten. Eine (zumindest partielle) syntaktische Analyse sollte dies erkennen und somit ”girl biscuits“ nicht als Terminus oder Indexterm auswählen.

Gegen eine syntaktische Analyse spricht allerdings, daß sie ressourcen- und rechenaufwendig ist: meist wird Information über syntaktische und semantische Eigenschaften von Wörtern (z.B. Valenz von Verben) benötigt, die in einem großen Lexikon abgespeichert wird, welches dann für jede Fachdomäne mit großem Aufwand entsprechend erweitert werden müßte.

Die Erstellung von Lexika und ihre fachspezifische Erweiterung ist aber gerade der Zweck der Terminologie-Extraktion, darf also keine Voraussetzung dafür sein. Damit erübrigt sich dieser Ansatz für meine Zwecke. Ein Chunking, welches mit flacheren Methoden arbeitet und somit einen Kompromiß zwischen POS-Filtern und vollständigem Parsen darstellt, wäre allerdings durchaus erwägenswert.

2.2 Sonstige Ansätze

2.2.1 Rein linguistische Verfahren

[Bourigault 1992]

Einer der ersten und richtungsweisenden rein linguistischen Ansätze zur Terminologie-Extraktion stammt von [Bourigault 1992]: sein System LEXTER wurde zur Pflege von Thesauri für die Firma *Electricité de France* entwickelt.

Bourigault geht davon aus, daß terminologische Einheiten Konzepte repräsentieren (in dem Sinne, wie in Abschnitt 1.1 erläutert), und zwar vollständig und eindeutig. Um Ambiguitäten zu vermeiden, müssen Termini also, laut Bourigault, eine ganz bestimmte Form haben, die sich für ihre

Extraktion ausnützen läßt: sie sind fast immer Nominalphrasen, die aus Nomina, Adjektiven und bestimmten Präpositionen (im Französischen sind das „de“ und „à“) bestehen dürfen, nicht aber aus sonstigen Wörtern wie Verben, Adverben, Artikeln etc.[1]

Um solche Nominalphrasen zu extrahieren, geht Bourigault in zwei Schritten vor:

1. oberflächliche syntaktische Analyse: ausgehend von einem POS-Tagging des Textes identifiziert das System maximale Nominalphrasen (NPs) mit Hilfe sogenannter “frontier markers“: Wörter oder Zeichen, die keinesfalls innerhalb eines Fachterminus auftreten und somit begrenzende Funktion haben, wie z.B. Satzzeichen, konjugierte Verben, Artikel und Pronomina.

2. diese maximalen NPs werden dann geparst, um terminologische Einheiten herauszufiltern. Dazu werden „handgemachte“ Regeln verwendet, die mit Hilfe eines umfangreichen Testkorpus auf ihre Zuverlässigkeit hin geprüft wurden. Ein Beispiel für eine solche Regel ist:

$$\text{noun1 adj prep det noun2 prep noun3} \rightarrow \begin{cases} \text{noun1 adj} \\ \text{noun2 prep noun3} \end{cases} \tag{2.6}$$

   Diese Regel zerlegt z.B. die maximale NP „disque dur de la station de travail“ in ihre Bestandteile ”disque dur“ und „station de travail“. Diese werden dann einem Terminologen zur Validierung vorgelegt.

Bourigaults System kombiniert also eine sehr primitive syntaktische Analyse des Textes (Auffinden von NPs mit Hilfe von „frontier markers“) mit einem POS-Filter für die erhaltenen maximalen NPs. Dies ist sicherlich sehr effizient, setzt aber ziemlich viel Handarbeit für das Erstellen der frontier markers und der Regeln für das Parsen voraus: Um sein System für das Englische oder Deutsche anzupassen, müßte man z.B. Vieles neu überlegen,

[1] Artikel können z.B. Ambiguitäten auslösen, wie Bourigault an dem Beispiel „écran d’un ordinateur portable“ illustriert: Weglassen des Artikels führt zur gebräuchlicheren — weil eindeutigeren — Version „écran d’ordinateur portable“

da in beiden Sprachen — anders als im Französischen — Modifikatoren in NPs dem Kopf der Phrase vorangestellt sein können (z.B. Adjektive). Sein Verfahren ist also leider sehr sprachabhängig.

[Arppe 1995]

Auch [Arppe 1995] extrahiert maximale Nominalphrasen aus Texten: er benutzt dazu einen Chunker, *NPtool*, und präsentiert *alle* in den maximalen NPs enthaltenen, korrekten Teil-Nominalphrasen, geordnet nach der Frequenz des grammatischen Kopfes.

Das sind natürlich viel zu viele, so daß er sich gezwungen sah, ein Kriterium zur Auslese zu finden: er untersuchte, welchen POS-Mustern die gefundenen Terme jeweils entsprachen und kam dabei unter anderem zu folgendem Ergebnis: ca. 80% aller Terme lassen sich durch eines der folgenden fünf Muster beschreiben (A steht hierbei für Adjektiv, N für Nomen, P für Präposition):

A N N N N N P N A N N

Diese Liste ist geordnet nach dem Recall des jeweiligen Musters, d.h. die meisten Terme entsprachen dem Muster A N usw. Interessant ist aber nicht nur der Recall eines Musters, sondern auch dessen Precision, d.h. die Frage danach, wieviele der mit seiner Hilfe extrahierten Wortgruppen tatsächlich Terme waren.

Generell gilt hierbei: kurze Muster (vor allem N und A N) haben zwar einen hohen Recall aber sehr geringe Precision, während lange Muster sehr genau sind, dafür aber kaum Recall haben. Bildet man den Mittelwert aus Precision und Recall, so schneiden die Muster N N, A N und A N N am besten ab.

Dies ist natürlich eine sehr interessante Betrachtung im Hinblick auf die Konstruktion eines Verfahrens, welches nur auf die flache Extraktion von POS-Mustern setzt: sie gibt Auskunft darüber, welche POS-Muster man extrahieren sollte und welche nicht. Eine genau Übersicht der von Arppe ermittelten Werte findet sich in Tabelle A.1 im Anhang.

Es gibt noch etliche weitere rein linguistische Verfahren, auf die aber an dieser Stelle nicht eingegangen werden soll, da sie entweder viel zusätzliche Information benötigen (wie z.B. vorher angelegte Listen domänenspezifischer Suffixe oder Basismorpheme) oder dem bisher Genannten nichts wesentlich Neues hinzufügen. Im Folgenden soll vielmehr auf eine Reihe möglicher Kombinationen von Wortstatistiken und linguistischen Methoden eingegangen werden.

2.2.2 Hybride Verfahren

[Daille 1994]

Einer der ersten richtungsweisenden (und vielzitierten) hybriden Ansätze zur Terminologie-Extraktion stammt von Daille, Gaussier und Langé (vgl. [Daille 1994]).Dort werden für das Englische nur die POS-Muster A N und N N betrachtet; diese werden als sogenannte „base MWUs“ (MWU steht für „multi-word unit“) extrahiert, um dann in eine Art kanonischer Form gebracht zu werden, die aus einem geordneten Paar von zwei Grundformen besteht.

In einem zweiten Teil der Untersuchung wird nach einem geeigneten statistischen Maß gesucht, um die Zusammengehörigkeit (Unithood) der gefundenen Paare herauszufinden und als Filter zu benutzen. Dabei wurden viele verschiedene Maße betrachtet und verglichen. Auch der Frage einer geeigneten Kombination von Maßen wurde mittels einer Hauptkomponentenanalyse nachgegangen.

Das Ergebnis ist mehr als erstaunlich: Es gibt laut [Daille 1994] *keine* geeignete Kombination von statistischen Maßzahlen. Vielmehr ist die beste Art, Zusammengehörigkeit und Fachspezifik einer Phrase zu beurteilen die Betrachtung der puren Frequenz des Paares im Text. Die beste statistische Prüfgröße für Unithood liefert nach den Ergebnissen von Daille der likelihood ratio Test (vgl. [Dunning 1993] und Abschnitt 3.2), er ist aber nur für niederfrequente Paare interessant.

Dieses Ergebnis sollte vielleicht nicht verallgemeinert werden (d.h. man darf es nur auf die Bewertung von Phrasen beziehen, die mittels POS-Mustern aus Texten extrahiert wurden). Die Untersuchungen wurden an

einem französischen Korpus von 240.000 Wörtern durchgeführt, es fragt sich also noch, wie repräsentativ die Ergebnisse für andere Sprachen sind. Insgesamt aber ist Dailles Resultat ein wichtiger Hinweis darauf, daß oft die einfachsten Methoden die besten sind.

[Justeson 1995]

Eine ganz ähnliche Idee verfolgt [Justeson 1995]. Hier findet sich eine theoretische Begründung für die von Daille gefundenen Resultate: Eine wichtige Eigenschaft terminologischer NPs ist — laut Justeson —, daß sie lexikalisch sind, d.h. daß ihre Bedeutung nicht komplett aus den Bedeutungen der Teile erschlossen werden kann und sie deshalb Aufnahme in ein Lexikon finden müssen.

Bei lexikalischen Einheiten ist aber formale Variation, insbesondere das Einfügen von Modifikatoren und Umstellung, sehr selten, da ihre Form ja im Lexikon festgeschrieben ist. Nicht-terminologische NPs sind dagegen auch meist nicht lexikalisch und treten somit — wenn sie überhaupt wiederholt werden — in verschiedenen Varianten auf. Daraus läßt sich ableiten, daß das *wiederholte* und *unvariierte* Auftreten einer NP ein Kriterium für Fachspezifik (*Termhood*) ist.

Aus diesen Überlegungen ist leicht ersichtlich, warum die pure Frequenz einer Phrase also ein gutes Maß für Termhood ist und man auf kompliziertere statistische Tests sowie auf kanonische Formen verzichten sollte. [Justeson 1995] akzeptiert einfach alle Phrasen mit einer gewissen Mindestfrequenz als Termini.

Problematisch bei Justesons Ansatz ist lediglich, daß sogenannte „Hintergrundterme“ nicht extrahiert werden, also solche, die als bekannt vorausgesetzt und daher oft nur einmal erwähnt werden. Ebenso muß die Mindestfrequenz dynamisch angepaßt werden, wenn man längere Texte untersucht, da dort auch nicht-terminologische NPs gelegentlich (unvariiert) wiederholt werden.

Auch die Struktur terminologischer NPs wird untersucht, wobei Justeson zu dem Ergebnis kommt, daß diese zu ca. 97% nur aus Adjektiven und Nomina bestehen und daß ihre durchschnittliche Länge (d.h. Anzahl der

Wörter) 1,91 beträgt. Dazu wurden Terme aus verschiedenen Fachlexika untersucht. In den meisten Domänen besteht die überwiegende Zahl der terminologischen NPs aus zwei Wörtern. Daher verwendet Justeson dieselben POS-Muster wie Daille (ergänzt um einige längere). Seine Betrachtungen über die Struktur von NPs gelten natürlich nur für das Englische.

[Frantzi 1996]

[Frantzi 1996] verzichtet weitgehend auf linguistische Methoden: hier wird nach häufig gemeinsam auftretenden Wortgruppen (Kollokationen) gesucht, wobei ein Hauptaugenmerk auf Verschachtelung liegt, d.h. auf der Frage, welche Teilstrings von Kollokationen selbst wieder interessante Wortgruppen sind.

Hierzu ein Beispiel: Die Phrase „New York Stock Exchange“ enthält drei Teilphrasen der Länge zwei: „New York“, „York Stock“ und „Stock Exchange“. Nur die erste und letzte dieser drei Wortgruppen sind aber selbst interessante Kollokationen. Das Maß *C-Value*, das in [Frantzi 1996] eingeführt wird, soll diese herausfiltern.

Dazu wird ausgenutzt, daß „New York“ und „Stock Exchange“ auch alleine oder als Teile anderer Phrasen auftreten. Der $C\text{-}Value(a)$ eines Kollokationskandidaten a berechnet sich wie folgt:

1. Falls a kein Teil längerer Kandidaten ist:

   $$C\text{-}Value(a) = (\|a\| - 1)\, n(a)$$

   $n(a)$ bezeichnet dabei die Frequenz von a, $\|a\|$ die Zahl der Wörter in a. Längere und häufige Kandidaten erhalten also einen hohen $C-Value$.

2. Falls a in mehreren längeren Kandidaten auftritt:

   $$C\text{-}Value(a) = (\|a\| - 1)\left(n(a) - \frac{t(a)}{c(a)}\right)$$

   $t(a)$ ist die Frequenz von a in längeren Kandidaten, $c(a)$ die Anzahl der Kandidaten, die a enthalten. Der *C-Value* wird hier also groß, wenn a oft alleine oder in vielen verschiedenen anderen Kandidaten auftritt.

   Insbesondere gilt $C\text{-}Value(a) = 0$, falls a nur als Teil *eines* längeren Kandidaten auftritt.

Je höher der *C-Value* eines Kandidaten, desto interessanter ist dieser im Sinne einer Kollokation.

Problematisch ist hierbei allerdings, daß sehr viele uninteressante Kollokationen, wie z.B. „said that“, extrahiert werden, die keine Nominalphrasen sind. Frantzi schlägt (wenn auch nur nebenbei) zur Behebung dieses Problems einen POS-Filter vor, geht damit also ein Stück in dieselbe Richtung wie die vorher genannten Ansätze. Interessant ist aber vor allem, wie Kollokationen, die nur als Teile längerer Kandidaten auftreten, eliminiert werden (z.B. „York Stock“).

Allen hybriden Verfahren gemeinsam ist also die Tatsache, daß sie auf komplexe statistische Tests verzichten und die Qualität von Phrasen einfach an ihrer puren Frequenz messen. Auch aufwendige linguistische Methoden (z.B. NP-Chunking oder Parsing) finden keine Verwendung, stattdessen werden einfache POS-Filter benutzt.

In [Justeson 1995] wird beschrieben, warum bei der Arbeit mit Fachterminologie trotz dieser primitiven Methoden gute Ergebnisse erzielt werden können: wichtige terminologische NPs haben eine vorhersagbare Struktur und treten häufig und unvariiert auf.

Man sollte also sorgfältig überlegen, ob und wann die Verwendung weitaus komplexerer Methoden lohnt, bzw. ob nicht manchmal die einfachste Methode die beste ist.

2.3 Ein Ansatz für das Deutsche

Alle bisher beschriebenen Verfahren sind entweder für Englisch oder Französisch entwickelt worden. Zwar sollten viele der dort gewonnenen Erkenntnisse auf das Deutsche übertragbar sein, doch es gibt einige Besonderheiten des Deutschen, die Probleme aufwerfen bzw. sich gut zusätzlich ausnutzen lassen.

Ein interessanter Beitrag zu diesem Thema findet sich in [Heid 1998]: er extrahiert Terminologie in mehreren Schritten und nutzt dabei besonders die morphologische Struktur des Deutschen (also die Existenz von Komposita

als Einworttermen). Dabei geht er wie folgt vor:

1. Extrahiere domänenspezifische Morpheme: Dazu gehören Basismorpheme, wie z.B. **fahr** oder **trieb**, aber auch Affixe wie *-ator* oder *-graph*. Um dies zu bewerkstelligen:

   - führe eine Differenzanalyse für Einwortterme durch. Das Ergebnis ist eine erste Liste von Fachtermini
   - zerlege Komposita und Derivate unter den gefundenen Einwortermen in ihre Bestandteile (Morpheme), z.B. „Ansaug - luft - temperatur“. Alle Morpheme, deren Frequenz über einem bestimmten Schwellenwert liegt, sind domänenspezifisch.

   Die so gewonnene Liste fachspezifischer Morpheme wird in den nächsten Schritten benutzt:

2. Extrahiere weitere Einwortterme, welche domänenspezifische Basismorpheme oder Affixe enthalten.

3. Extrahiere Phrasen nach bestimmten POS-Mustern und filtere diese unter Benutzung der Einwortterme aus der Differenzanalyse.

Die POS-Muster für das Deutsche sind natürlich etwas anders als für Englisch. Heid verwendet die Muster

- N Det N^{Gen} („Ablauf der Frist“)
- N P N („Prüfkupplung für Diagnose“) und
- A N („verstellbarer Außenspiegel“).

Akzeptiert werden aber nur solche Phrasen, bei denen mindestens eins der enthaltenen Wörter schon durch die Differenzanalyse als Terminus festgestellt wurde.

Das muß so streng gehandhabt werden, weil Mehrwortgruppen im Deutschen insgesamt seltener sind als im Englischen und daher mehr Rauschen entsteht, wenn man nur nach POS-Mustern sucht.

Insbesondere bei Adjektiv-Nomen Mustern ist es wichtig, daß das Adjektiv „Termstatus“ hat, da viele Adjektive (wie z.B. „möglich“, „folgend“ etc.)

nur in uninteressanten NPs auftreten. In der Praxis benutzte Heid daher eine „Stopliste“ solcher Adjektive, um Phrasen auszuschließen, welche diese enthielten.

Heids gesamter Ansatz scheint mir unter anderem deshalb sehr interessant, weil er sich gut in ein System integrieren läßt, welches auf Relevance Feedback aufbaut: In einem ersten Schritt wird eine Differenzanalyse durchgeführt, der Anwender gibt ein Urteil über die gefundenen Einwortterme ab; daraus lassen sich Basismorpheme herausarbeiten, die für den Benutzer relevant waren. Mit deren Hilfe kann man schließlich weitere Terme auffinden.

Der zweite Teil dieses Vorgehens wird auch *Bootstrapping* genannt: eine „Startmenge“ an gesichertem Wissen wird schrittweise erweitert, indem die Eigenschaften der Elemente der Startmenge in eine Regelmenge umgesetzt werden, die es ermöglicht, weitere Terme zu finden.

Ich werde daher die Architektur dieses Ansatzes grob übernehmen und in meinem eigenen Verfahren verarbeiten. Dieses soll nun im Folgenden dargestellt werden.

Kapitel 3

Ein eigenes Verfahren

Ich möchte nun ein von mir selbst entwickeltes Verfahren beschreiben, welches ich zunächst nur für das Englische implementiert habe. Kapitel 5 beschreibt Anpassungen für das Deutsche.

Die Architektur des Systems soll durch Abbildung 3.1 veranschaulicht werden.

Die **Vorverarbeitung** des Textes umfaßt zunächst Filteroperationen (z.B. das Wegfiltern von HTML-Tags), dann:

- Segmentieren des Textes in Sätze und Wörter,
- Entfernen von Stopwörtern,
- Grundformenreduktion,
- Bestimmung der Frequenz von Grundformen und Buchstabentrigrammen im Text,
- Aufruf des POS-Taggers, Zwischenspeichern seiner Ausgabe in einer temporären Datei,
- Anlegen einer inversen Liste, d.h. das System merkt sich zu jedem Wort, in welchen Sätzen es aufgetreten ist.

Die Segmentierung der Sätze und Wörter, die Grundformenreduktion und das POS-Tagging werden in Abschnitt 3.1 näher beschrieben.

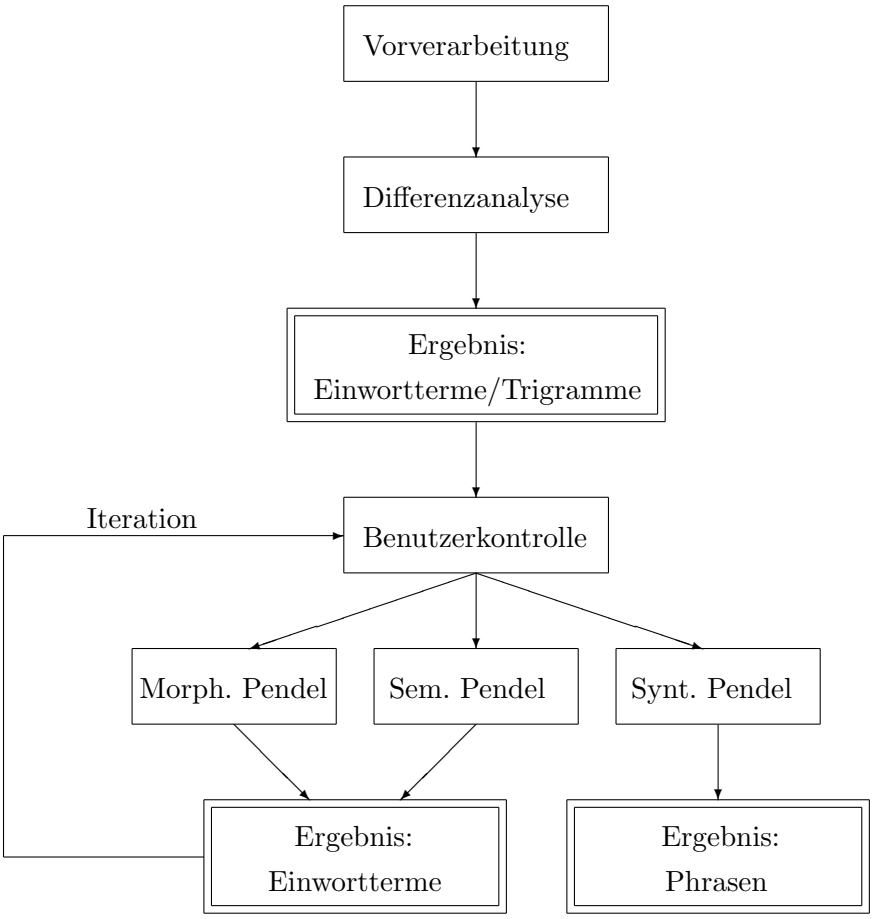

Abbildung 3.1: Architektur der Extraktionssoftware

Bei der nachfolgenden **Differenzanalyse** werden die Frequenzen der Grundformen und der Buchstabentrigramme mit den entsprechenden Frequenzen im Referenzkorpus verglichen. Eine genauere Beschreibung des Vorgehens findet sich in Abschnitt 3.2.

Die Ergebnisse der Differenzanalyse werden dem Anwender zur Beurteilung vorgelegt. Dieser wählt nun aus, welche davon ihm als relevante Terminologie erscheinen bzw. welche Trigramme er tatsächlich für domänenspezifisch hält. Nur die vom Benutzer ausgewählten Terme und Trigramme finden später Eingang in die Ergebnismenge *und*: nur sie und ihre Eigenschaften werden berücksichtigt, um nach neuen Termen zu suchen.

Nun wird mittels der in Abschnitt 3.3 beschriebenen **„Pendel“** nach

weiteren Termen gesucht; diese werden wieder dem Benutzer vorgelegt, erhalten von diesem Termstatus zugewiesen oder nicht usw. Der Prozeß wird solange iteriert, bis keine neuen Terme mehr hinzukommen. Man kann die verschiedenen Pendel dabei nach Belieben an- und ausschalten.

Der ganze Prozeß ist halbautomatisch, da der Benutzer nach jedem Schritt verifizieren muß, welche der gefundenen Kandidaten tatsächlich Terme in seinem Sinne sind. Seine Auswahl hat aber natürlich Einfluß auf den nächsten Iterationsschritt, so daß man tatsächlich von Relevance Feedback sprechen kann.

In den folgenden Abschnitten möchte ich nun die einzelnen Komponenten des Systems genauer beschreiben.

3.1 Vorverarbeitung des Textes

3.1.1 Wort- und Satzsegmentierung

Das erste Problem bei der Aufbereitung eines Fachtextes besteht darin, die Grenzen zwischen Sätzen und Wörtern herauszufinden. Man mag zunächst versucht sein, diese Aufgabe als einfach anzusehen, da Sätze meist mit einem Punkt, Ausrufe- oder Fragezeichen enden und Wörter meist durch Leerzeichen getrennt sind.

In der Praxis gestaltet sich die Segmentierung aber schwieriger (vgl. [Manning 2002], Abschnitt 4.2): Zunächst muß man sich klarmachen, daß nach einem Wort auch ein beliebiges Satzzeichen stehen kann. Man definiert ein Wort also als eine Folge von mindestens zwei Buchstaben (Satzzeichen ausgenommen, Bindestriche eingeschlossen), begrenzt durch ein Leerzeichen oder ein Satzzeichen.

Dabei stößt man aber auf das nächste Problem: Abkürzungen wie *etc.* enden mit einem Punkt, der Teil des Wortes ist. Hier hilft eine Liste von Abkürzungen, bei denen der Punkt als wortinternes Zeichen erlaubt ist. Es gibt noch einige andere Probleme bei der Wortsegmentierung — z.B. die Frage, ob ein Apostroph ein Wort beendet oder nicht — die ich aber bei der Implementierung nicht beachtet habe.

Weiter läßt sich sagen, daß es in „realen“ Texten eine große Zahl von

Wörtern gibt, die seltsame Zeichen enthalten (z.B. „Micro$oft“). Verbietet man Zeichen wie $ innerhalb eines Wortes, so verzichtet man auf diese Wörter, aber eben auch auf viele Zeichenketten, die wirklich keine Wörter sind (z.B. „$22.50“).

Ich habe mich deshalb dazu entschlossen, recht strenge Kriterien für Wörter zu verwenden, um das Rauschen gering zu halten. So darf ein Wort bei meiner Implementierung nur die Zeichen a-z, A-Z, ä, ö, ü, ß, den Punkt und einen Bindestrich enthalten; insbesondere Ziffern sind innerhalb von Wörtern verboten.

Nicht viel einfacher gestaltet sich die Segmentierung des Textes in Sätze. Nimmt man an, daß ein Satz durch eines der Zeichen .:!? beendet wird, so hat man wieder das Problem der Abkürzungen: diese enden mit einem Punkt, der aber nicht unbedingt ein Satzende signalisiert (z.B. „z.B.“). Hier kann wieder eine Abkürzungsliste helfen; allerdings gibt es auch Abkürzungen (wie z.B. „etc.“), die durchaus am Satzende stehen können.

Eine gute Heuristik ist in diesem Fall (zumindest für das Englische, wo Großbuchstaben selten sind) die Forderung, daß nach einem Satz ein Leerzeichen und danach ein Großbuchstabe folgen muß.

Ein weiteres Problem ist die direkte Rede innerhalb eines Satzes:

’Sie sollten einen Regenschirm mitnehmen. Es regnet’, sagte er.

Die bisher entwickelte Heuristik liefert zwei Sätze: „’Sie sollten einen Regenschirm mitnehmen.“ und „Es regnet’, sagte er.“ Das ist offensichtlich falsch. Allerdings ist nicht ganz klar, wie man richtig segmentieren sollte, denn einerseits ist das Ganze ein Satz, andererseits enthält dieser wieder zwei vollständige Sätze. Würde man das Ganze als nur einen Satz behandeln, so könnten im Falle längerer direkter Rede sehr lange Sätze entstehen.

Ein ähnliches Problem — welches allerdings mit der Textformatierung zusammenhängt — sind Tabellen: hat man die Tags aus einer HTML-Datei entfernt, so erscheinen die Inhalte von Tabellen als ein Satz, da zwischen den einzelnen Zellen der Tabelle keine Punkte stehen. Das kann zu sehr langen Sätzen führen.

Mir ist es nicht gelungen, alle diese Probleme absolut befriedigend zu

lösen, d.h. es tauchen im Verlauf der Extraktion ab und zu Wörter und Sätze auf, die eigentlich keine sind. Es gibt Ansätze, die versuchen, verschiedene Lernverfahren für eine perfektere und weniger sprachabhängige Lösung der oben beschriebenen Probleme zu entwickeln (vgl. [Manning 2002], S. 135).

Ich habe mich aber dennoch für eine eher einfache Heuristik entschieden, einerseits aus Effizienzgründen, andererseits, weil ich denke, daß die Qualität der Segmentierung für meine Zwecke ausreicht (Sätze sollen nur als Beispielsätze dienen).

3.1.2 Grundformenreduktion

Hat man nun den Text in Sätze und Wörter zerlegt, so muß als nächstes die Flexion der Wörter neutralisiert werden, da ich mich dafür interessiere, wie häufig ein *Wort* in einem Text auftritt, nicht aber dafür, wie häufig die einzelnen *Wortformen*, d.h. die flektierten Formen des Wortes, vorkommen.

Die Reduktion von Wörtern auf ihre Grundform (auch Lemmatisierung genannt) soll dabei ohne Verwendung eines Lexikons stattfinden, da es ja bei der Terminologie-Extraktion oft darum geht, Wörter zu finden, die eben noch in keinem Lexikon stehen.

Nun hat das Englische nur sehr wenige Flexionssuffixe („-s“, „-ed“ und „-ing“) und es treten auch bei der Flexion nur wenige Allomorphe auf. Das bedeutet, daß man Wörter recht leicht auf ihre Grundform reduzieren kann, indem man einfach die gerade genannten Flexive hinten von Wörtern abschneidet. Natürlich muß man ein paar Ausnahmen beachten (und natürlich wird man nicht alle beachten *können*).

Beim Abschneiden von „-s“ gibt es z.B. zwei wichtige Ausnahmen, die man beachten sollte:

- Endet das Wort auf „-ss“, so soll das letzte „s“ nicht abgeschnitten werden (Bsp.: „caress“)
- Endet das Wort auf „ies“, so sollen diese drei Buchstaben im Ganzen abgeschnitten und durch „y“ ersetzt werden (Bsp.: ponies → pony)

Ein weitverbreiteter Algorithmus zur Grundformenreduktion im Englischen ist der *Porter-Algorithmus* (vgl. [Porter 1980]). Dieser behandelt etli-

che wichtige Ausnahmen (aber natürlich nicht alle; er ist also nicht fehlerfrei!) mit Hilfe eines insgesamt recht übersichtlichen Regelwerks. Allerdings neutralisiert er nicht nur Flexion, sondern auch Derivation; man kann jedoch recht einfach den Teil isolieren, der sich mit Flexion befaßt. Ich habe also eine leicht abgewandelte Version des Porter-Algorithmus verwendet.

3.1.3 POS-Tagging

Für das POS-Tagging verwendete ich den Tagger *TnT* (TnT steht für „Trigrams'n'Tags“) von Thorsten Brants (vgl. [Brants 2000]). Dieser basiert auf Markov-Modellen zweiter Ordnung — d.h. es werden immer Trigramme von Wörtern bzw. Tags betrachtet — und erwies sich im Vergleich zu einem regelbasierten Tagger (vgl. [Brill 1992]) als wesentlich zuverlässiger.

Die grundlegende Idee eines Taggings mit Markov-Modellen ist es, einer Folge von Wörtern $w_1 \dots w_n$ die *wahrscheinlichste* Folge von POS-Tags $t_1 \dots t_n$ zuzuordnen. Dabei nutzt man aus, daß es gewisse Beschränkungen der Syntax gibt, d.h. daß manche Tag-Folgen wahrscheinlicher sind als andere. Zum Beispiel folgt auf einen Artikel eher ein Nomen als ein Verb.

Es soll also zu einer gegebenen Wortfolge $w_1 \dots w_n$ berechnet werden:

$$\text{argmax}_{t_1 \dots t_n} P(t_1 \dots t_n | w_1 \dots w_n) \tag{3.1}$$

Benutzt man nun die Bayes'sche Regel, so erhält man:

$$\text{argmax}_{t_1 \dots t_n} \frac{P(w_1 \dots w_n | t_1 \dots t_n) P(t_1 \dots t_n)}{P(w_1 \dots w_n)} \tag{3.2}$$

Den Nenner dieses Ausdrucks kann man ignorieren, da er nicht von der Tag-Folge abhängt. Für den Zähler macht man folgende Annahmen:

- Die Wahrscheinlichkeit eines Tags hängt immer nur von den zwei vorangehenden Tags ab, d.h. $P(t_n | t_1 \dots t_n) = P(t_n | t_{n-2}, t_{n-1})$
- Die Wahrscheinlichkeit eines Wortes hängt nicht von den umgebenden Wörtern ab, d.h. $P(w_i | w_1, \dots, w_n) = P(w_i)$ (was offensichtlich grob vereinfachend ist, da z.B. das Wort „New“ wahrscheinlicher vor „York“ auftritt als in beliebigen anderen Kontexten).

- Die Wahrscheinlichkeit eines Wortes hängt nur von seinem eigenen Tag (also nicht von umliegenden Tags) ab, d.h. $P(w_i|t_1 \dots t_n) = P(w_i|t_i)$

Mit diesen Annahmen vereinfacht sich Formel (3.2) zu:

$$argmax_{t_1 \dots t_n}(\prod_{i=1}^{n} P(t_i|t_{i-1}, t_{i-2})P(w_i|t_i)) \tag{3.3}$$

Die in dieser Formel vorkommenden Wahrscheinlichkeiten lassen sich anhand eines von Hand getaggten Trainingskorpus einfach abschätzen:

$$P(t_i|t_{i-1}, t_{i-2}) = \frac{f(t_{i-2}, t_{i-1}, t_i)}{f(t_{i-2}, t_{i-1})} \tag{3.4}$$

$$P(w_i|t_i) = \frac{f(w_i, t_i)}{f(t_i)} \tag{3.5}$$

f bezeichnet dabei jeweils die Frequenz. Man nennt diese Schätzungen *maximum likelihood estimates*, da die Wahrscheinlichkeiten einfach aus den Häufigkeiten im Trainigskorpus vorhergesagt werden.

Oft wird noch ein sogenanntes *Smoothing* durchgeführt, d.h. die Trigrammwahrscheinlichkeiten $P(t_i|t_{i-1}, t_{i-2})$ werden ersetzt durch $\lambda_1 P(t_i) + \lambda_2 P(t_i|t_{i-1}) + \lambda_3 P(t_i|t_{i-1}, t_{i-2})$ mit $\lambda_1 + \lambda_2 + \lambda_3 = 1$. Dies ist notwendig, da manche Trigramme im Korpus gar nicht auftreten; für diese ist $P(t_i|t_{i-1}, t_{i-2}) = 0$, so daß der ganze Ausdruck aus Formel (3.3) verschwindet, obwohl womöglich nur dieses eine Trigramm nie aufgetreten ist. Dies kann durch Smoothing vermieden werden.

Das eigentliche POS-Tagging läuft nun folgendermaßen ab: Zunächst wird ein Lexikon benutzt, in dem für viele, aber sicher nicht alle Wörter des Textes festgehalten ist, welche POS-Tags sie haben *können*. Es ist wichtig, sich klarzumachen, daß manche Wörter durchaus mit mehreren POS-Tags vorkommen können (z.B. kann das englische Wort „set“ sowohl Nomen als auch Verb sein).

Für alle unbekannten Wörter wird ein POS-Tag geraten. Hierbei kann die Endung des Wortes hilfreich sein: englische Wörter, die auf „-able“ enden, sind z.B. meist Adjektive. Gibt die Endung keinerlei Hinweis auf die Wortart, so wird ein beliebiges Tag (z.B. Nomen) zugewiesen.

Nun hat man also für manche Wörter mehrere, für andere Wörter zweifelhafte Tags. Im nächsten Schritt kann man obige Formeln verwenden, um aus allen möglichen Tag-Folgen die wahrscheinlichste herauszusuchen.

Der tatsächlich verwendete Algorithmus ist um einiges komplexer, da Optimierungen vorgenommen werden, so daß nicht für alle Tag-Folgen die Wahrscheinlichkeit explizit berechnet werden muß. Das Produkt des POS-Taggings ist jedenfalls eine *eindeutige* Zuordnung von POS-Tags zu Wörtern.

Nach diesen vorbereitenden Schritten kann jetzt die eigentliche Terminologie-Extraktion beginnen.

3.2 Schritt 1: Differenzanalyse

Der erste Schritt meiner Terminologie-Extraktion besteht aus einer Differenzanalyse, d.h. der Suche nach Wörtern, deren Frequenz im Fachtext statistisch signifikant abweicht von der Frequenz, die aufgrund des Vergleichskorpus vorhergesagt wird.

Zunächst mußte ich dazu einen geeigneten *statistischen Test* auswählen und diesen implementieren (tatsächlich habe ich mehrere implementiert, um sie später vergleichen zu können). Dasselbe sollte dann für *Buchstabentrigramme* geschehen. Schließlich erwies es sich als sinnvoll, *Eigennamen* unter den Termen zu identifizieren und getrennt auszugeben.

3.2.1 Suche nach einer statistischen Prüfgröße

Wie im Abschnitt 1.3 angekündigt, soll nun nach einer geeigneten statistischen Prüfgröße P für die Differenzanalyse gesucht werden. Zur Erinnerung: diese soll möglichst groß werden, wenn die Frequenz eines Wortes statistisch signifikant über der Frequenz liegt, die aufgrund seiner Frequenz im Vergleichskorpus zu erwarten gewesen wäre.

Für jedes Wort w eines Fachtextes der Länge n stellt man sich dabei das folgende statistische Experiment vor: man vergleicht w mit einem Wort x des Textes; das Ergebnis des Experiments ist positiv, wenn $x = w$, sonst negativ. Diesen Test wiederholt man nun für jedes Wort des Fachtextes, also n-mal, und interessiert sich nun für die Wahrscheinlichkeit von k positiven Ergebnissen. D.h. man will wissen, wie wahrscheinlich es ist, daß w genau k-mal im Text auftritt.

Diese Wahrscheinlichkeit hängt natürlich davon ab, wie wahrscheinlich es ist, bei einem einzelnen Experiment ein positives Ergebnis zu erzielen,

oder anders gesagt: wie hoch die Auftretenswahrscheinlichkeit p des Wortes w in einem beliebigen Text ist.

Den exakten Wert von p könnte man nur berechnen, wenn man alle Texte dieser Welt zur Verfügung hätte. Da dies nicht der Fall ist, begnügt man sich damit, p anhand des Referenzkorpus zu schätzen. Dazu nähert man p einfach durch die relative Häufigkeit $\frac{f(w)}{N}$ des Wortes w im Vergleichskorpus an (maximum likelihood estimate).

Die Idee des statistischen Testens ist es nun, eine tatsächlich beobachtete Verteilung (z.B. eines Wortes w in einem Fachtext) mit der Verteilung in einer (z.B. anhand eines Referenzkorpus geschätzten) Grundgesamtheit zu vergleichen.

Dabei geht man in fast allen Fällen von einer sogenannten *Nullhypothese* aus, welche besagt, daß die beobachtete Verteilung der Verteilung in der Grundgesamtheit entspricht. Das heißt: die Wahrscheinlichkeit, das Wort w im Fachtext anzutreffen, ist gleich der aus dem Referenzkorpus geschätzten Wahrscheinlichkeit p.

Fast alle statistischen Tests bestehen nun darin, eine Prüfgröße zu berechnen, welche umso größer wird, je mehr das tatsächlich beobachtete Ergebnis von der Nullhypothese abweicht.

Die verschiedenen Tests unterscheiden sich zum einen in der Art, wie die Verteilung von Wörtern in Texten modelliert wird (Normal-, Binomial- oder Poissonverteilung) und in der verwendeten Prüfgröße (wobei das eine eng mit dem anderen zusammenhängt).

Der likelihood-ratio-Test

In gewisser Weise gleicht das oben beschriebene Experiment dem Werfen einer Münze (Kopf = positives, Zahl = negatives Ergebnis). Beide Experimente sind aber nur unter zwei Voraussetzungen identisch:

1. Jedes Experiment ist unabhängig von allen vorangehenden. Dies ist für Wörter nicht ganz korrekt, z.B. ist für w = „Energie“ die Wahrscheinlichkeit erhöht, wenn das vorangegangene Wort „kinetische“ war. Die Abhängigkeit zwischen beiden Wörtern verliert sich aber schon im übernächsten Experiment weitgehend.

2. Die Wahrscheinlichkeit p eines positiven Ergebnisses ist bei jedem Test der Serie gleich. Das hieße für Wörter insbesondere: das Wort w ist an jeder Stelle des Textes gleich wahrscheinlich. Das stimmt nicht ganz, einerseits wegen der oben besprochenen Abhängigkeiten, aber auch, da Themenwechsel innerhalb von Dokumenten und Dokumentenkollektionen eine wechselnde Wortwahl bedingen.

Beide Annahmen sind also nicht ganz erfüllt; man kann aber davon ausgehen, daß die Effekte der Abweichungen sich über den Text hinweg ausgleichen (hierfür gibt es keinen theoretischen Beweis, die Ergebnisse müssen mir Recht geben).

Das Werfen einer Münze und — nach unseren Näherungen — auch das Vergleichen von Wörtern mit dem Wort w wird durch die Binomialverteilung beschrieben. Das heißt: ist p die Wahrscheinlichkeit, bei einem Vergleich ein positives Ergebnis zu erzielen, so ist die Wahrscheinlichkeit, bei n Versuchen k positive Ergebnisse zu erzielen, gegeben durch:

$$p(n,k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.6}$$

Die Binomialverteilung hat einen Erwartungswert von np und eine Varianz von $np\,(1-p)$.

Oftmals wird die Binomialverteilung durch eine Normalverteilung angenähert. [Dunning 1993] kritisiert dies heftig und zeigt auf, daß für seltene Ereignisse (d.h. seltene *Wörter* in unserem Fall) die Normalverteilung signifikant von der Binomialverteilung abweicht.

In der natürlichen Sprache gibt es aber viele seltene Wörter: aus dem in Abschnitt 1.3 eingeführten Zipfschen Gesetz folgt auch, daß die Hälfte aller Wörter eines beliebigen Textes nur einmal in diesem auftreten!

[Dunning 1993] verwendet darum einen sogenannten *likelihood-ratio*-Test, der direkt auf der Binomialverteilung aufbaut. Dabei verwendet man die *likelihood-Funktion* H:

$$H(p_1, p_2; k_1, n_1, k_2, n_2) = \prod_{i=1,2} \binom{n_i}{k_i} p_i^{k_i} (1-p_i)^{n_i - k_i} \tag{3.7}$$

H beschreibt die Wahrscheinlichkeit, ein Wort w im Fachtext der Länge n_1 k_1-mal und im Vergleichskorpus der Länge n_2 k_2-mal zu sehen unter der

Voraussetzung, daß die Auftretenswahrscheinlichkeit im Fachtext durch p_1 und die im Korpus durch p_2 gegeben ist.

Die Nullhypothese lautet: $p_1 = p_2 = p$, d.h. w ist in beiden Texten gleichwahrscheinlich. Die likelihood ratio λ ist nun der Quotient zweier Maxima: dem maximalen Wert der likelihood-Funktion H auf dem Teilraum Ω_0, der durch die Nullhypothese gegeben ist, geteilt durch das Maximum von H auf dem gesamten Ereignisraum Ω:

$$\lambda = \frac{\max_{\omega \in \Omega_0} H(\omega; k)}{\max_{\omega \in \Omega} H(\omega; k)} \tag{3.8}$$

In unserem Fall ist dies:

$$\lambda = \frac{\max_p H(p, p; k_1, n_1, k_2, n_2)}{\max_{p_1, p_2} H(p_1, p_2; k_1, n_1, k_2, n_2)} \tag{3.9}$$

Die Maxima werden erreicht durch die *maximum likelihood estimates* $p_1 = \frac{k_1}{n_1}$, $p_2 = \frac{k_2}{n_2}$ und $p = \frac{k_1 + k_2}{n_1 + n_2}$. Nach Einsetzen und Umformen erhält man die eigentliche Prüfgröße $-2 \log \lambda$:

$$\begin{aligned} -2 \log \lambda = \; & 2[\log L(p_1, k_1, n_1) + \log L(p_2, k_2, n_2) \\ & - \log L(p, k_1, n_1) - \log L(p, k_2, n_2)] \end{aligned} \tag{3.10}$$

wobei $L(p, k, n) = p^k (1 - p)^{n-k}$ ist. Die Differenzanalyse besteht in diesem Fall also darin, alle Wörter, für die $-2 \log \lambda$ groß genug ist (und die mit einer gewissen Mindestfrequenz auftreten), herauszufiltern.

Das Poisson-Maß

Die Poisson-Verteilung ist eine bessere Näherung an die Binomialverteilung als die Normalverteilung. Für das Experiment, ein Wort w nacheinander mit allen n Wörtern eines Fachtextes zu vergleichen, besagt die Poisson-Verteilung, daß man dabei mit Wahrscheinlichkeit

$$p(n, k) = \frac{1}{k!} \lambda^k e^{-\lambda}, \quad \lambda = np \tag{3.11}$$

k Auftreten des Wortes w beobachtet. Dies gilt aber wieder nur unter der Nullhypothese, daß das Wort im Fachtext mit gleicher Wahrscheinlickeit

auftritt wie im Referenzkorpus. Die Korpuswahrscheinlichkeit p wird wieder durch die relative Häufigkeit von w im Korpus abgeschätzt. Die Poissonverteilung hat den Erwartungswert $\lambda = np$.

Eine an der Universität Leipzig zur Berechnung von Kollokationen eingesetzte statistische Prüfgröße hat ebendiese Poissonverteilung als Grundlage (vgl. [Quasthoff 2002]).

Die Formel für Kollokationen läßt sich aber leicht für eine Differenzanalyse anpassen: hat man das Wort w k-mal im Text beobachtet, so berechnet man die Wahrscheinlichkeit für *mindestens* k Auftreten mit Hilfe von Formel 3.11. Der negative Logarithmus dieser Zahl ist ein Maß für das „Erstaunen“, das Wort so oft gesehen zu haben unter Annahme des Erwartungswertes λ.

Die tatsächliche Formel enthält noch einen Normalisierungsfaktor:

$$sig(w) = -\frac{1}{\log n} \log \sum_{l=k}^{\infty} \frac{1}{l!} \lambda^l e^{-\lambda} \tag{3.12}$$

Mit einigen Annahmen an die Größenordnung von λ und etlichen Umformungen gelangt man zu der vereinfachten Formel:

$$sig(w) = \frac{k(\log k - \log \lambda - 1)}{\log n} \tag{3.13}$$

Diese kann aber nur dann zum Einsatz kommen, wenn die Bedingungen an λ erfüllt sind.

In [Quasthoff 2002] findet man auch einen Vergleich zwischen diesem Maß und dem von Dunning vorgeschlagenen likelihood-ratio-Test. Dabei wird klar, daß beide Formeln sich nur sehr geringfügig voneinander unterscheiden (zumindest wenn Formel 3.13 für das Poisson-Maß verwendet wird). Ich habe beide Prüfgrößen implementiert und festgestellt, daß auch die Ergebnisse fast ausnahmslos übereinstimmen.

Der Häufigkeitsquotient

Am Anfang meiner Experimente mit Differenzanalysen benutzte ich — so wie [Damerau 1993] — ein sehr einfaches Maß, den Häufigkeitsquotienten: die relative Häufigkeit des Worts w im Fachtext, geteilt durch seine relative Häufigkeit im Referenzkorpus (diese soll hier wieder p heißen).

$$hq(w) = \frac{k}{n} \frac{1}{p} \tag{3.14}$$

k bezeichnet dabei wieder die Frequenz von w im Fachtext, n dessen Länge in Wörtern.

Dieses Maß ist kein richtiger statistischer Test, sondern eher eine einfache Heuristik. Es kann aber als eine Art likelihood ratio interpretiert werden (vgl. [Manning 2002], S. 175), allerdings ohne zugrundeliegende Binomialverteilung.

Später (siehe Kapitel 4) wird daher Dunnings These zu prüfen sein, welche besagt: Verfahren, die auf Normalverteilung aufbauen (oder — wie der Häufigkeitsquotient — ganz auf eine statistische Modellierung verzichten) neigen zur Überbewertung seltener Terme.

Das liegt daran, daß Binomial- und Poissonverteilung gegenüber der Normalverteilung leicht „nach links verschoben“ sind, d.h.: das einmalige Eintreten eines in der Grundgesamtheit extrem seltenen Ereignisses ruft bei Verwendung einer Normalverteilung wesentlich mehr (nämlich zuviel) „Erstaunen“ hervor als bei einer Binomialverteilung.

[Damerau 1993] bezweifelt aber, daß man die Verteilung von Wörtern in Texten mittels Binomial- oder Poissonverteilung beschreiben kann: wie wir gesehen haben, wird dabei vorausgesetzt, daß das Wort w an jeder Stelle des Textes gleich wahrscheinlich ist. Damerau setzt dem entgegen, daß innerhalb eines jeden Textes Wörter dazu neigen, innerhalb kleiner Regionen (z.B. Absätze) zu clustern, so daß diese Voraussetzung nicht erfüllt ist.

Aufgrund der vielen widerstreitenden Meinungen darüber, welches Maß das beste ist, habe ich mich entschlossen, alle oben beschriebenen Maße zu implementieren. In Kapitel 4 möchte ich dann die Theorie ein wenig beiseite schieben und versuchen herauszufinden, welches Maß in der praktischen Anwendung am besten abschneidet.

3.2.2 Einwortterme

Wie in Abschnitt 1.3 angedeutet, lassen sich Einwortterme innerhalb eines Fachtextes mittels einer Differenzanalyse bestimmen, indem man alle Wörter auswählt,

- deren Frequenz oberhalb eines gewissen Mindestwertes liegt und

- für die der Wert einer der oben beschriebenen Prüfgrößen einen Schwellenwert überschreitet.

An dieser Stelle möchte ich nun auf die Auswahl eines geeigneten Referenzkorpus eingehen: Dieser sollte erstens groß sein und zweitens möglichst alle Gebiete abdecken, über die „jemals geschrieben wurde“. Dabei kommt es auch noch auf die richtige Mischung an: „wichtige“ Themen sollen stärker vertreten sein als „unwichtige“. Die Frage, was ein „wichtiges“ Thema ist, kann dabei aber von niemandem eindeutig beantwortet werden und ist eine Wissenschaft für sich. Es wird also kein vollkommen repräsentatives Korpus geben.

Ich habe ein an der Universität Leipzig verfügbares Korpus verwendet, welches aus einigen Jahrgängen einer Computerzeitschrift (*IEEE Transactions on Computers*), des Wall Street Journal, der Financial Times und weiteren Zeitungsartikeln zusammengestellt ist. Themen wie Wirtschaft, Politik und Informatik sind daher stark, andere dagegen weniger vertreten.

Insgesamt kann man wohl davon ausgehen, daß Zeitungsartikel eine sehr breite Auswahl an Themen bieten; das Problem liegt hauptsächlich darin, daß einige wenige Themen wie Wirtschaft und Informatik extrem überrepräsentiert sind.

Es ist daher zu erwarten, daß man bei der Analyse eines informatischen Fachtextes Probleme bekommt: informatikspezifische Fachbegriffe verursachen aufgrund des „IT-lastigen“ Referenzkorpus weniger „Erstaunen“. Bedenkt man dies nicht, so kann man wohl manchmal unangenehme Überraschungen erleben.

Beginnt man nun mit der Implementierung, so wird man sehr schnell feststellen, daß jedes noch so große und repräsentative Referenzkorpus in gewisser Weise unvollständig ist: manche Wörter des Fachtextes sind nicht im Vergleichskorpus enthalten.

Es stellt sich nun die Frage, was mit diesen Termen geschehen soll. Hier hilft wieder eine sehr einfache Heuristik: Wörter, die im Referenzkorpus nicht enthalten sind, sind meistens Fachtermini, wenn sie mit einer gewissen Mindestfrequenz auftreten (wenn sie nur einmal auftreten, handelt es sich dagegen meist um Rechtschreibfehler).

Das Nicht-Enthaltensein kann man als einen unendlich großen P-Wert interpretieren; die Mindestfrequenz verhindert die Überbewertung dieses seltenen Ereignisses. Die Praxis zeigt, daß diese Annahmen gerechtfertigt sind: nicht-enthaltene Wörter sind meistens wirklich Fachtermini.

Ist also die Differenzanalyse für alle Wörter des Textes abgeschlossen, so werden die gefundenen Terme dem Anwender zur Begutachtung vorgelegt. Als zusätzliche Information werden angezeigt:

- die Frequenz des Terms im Fachtext,
- der Wert der statistischen Prüfgröße P für diesen Term, sowie
- zwei Beispielsätze, die den Term enthalten

Der Anwender wählt nun aufgrund all dieser Informationen die Wörter aus, denen er Termstatus zuweisen möchte, und diese fließen in den weiteren Ablauf des Verfahrens ein.

3.2.3 Buchstabentrigramme

Dem Ansatz von [Cohen 1995] folgend, habe ich eine Differenzanalyse auch für Buchstabentrigramme implementiert, um domänenspezifische Buchstabenfolgen herauszufinden und mit deren Hilfe weitere Einwortterme zu identifizieren. Dabei konzentrierte ich mich zunächst auf Suffixe von Wörtern, stellte aber fest, daß sich die Ausbeute (bei gleichbleibender Präzision) erhöht, wenn man Trigramme an beliebigen Stellen im Wort betrachtet.

Ich ging also genauso vor wie bei Wörtern: die Frequenz aller Buchstabentrigramme wird sowohl für die Fachtexte als auch für das Referenzkorpus bestimmt und es kommen die gleichen statistischen Tests zur Anwendung wie für Wörter.

Die Ergebnisse werden im gleichen Format ausgegeben wie die Einwortterme. Allerdings wird nicht die Frequenz des Trigramms im Text angezeigt, sondern die Anzahl *verschiedener* Wörter, in denen das Trigramm vorkommt. Statt der Beispielsätze werden außerdem nur die ersten drei dieser Wörter aufgeführt.

Zwar ist die Grundlage für die Berechnung der Prüfgröße die tatsächliche Frequenz des Trigramms im Text, doch für den Anwender ist die Anzahl *ver-*

schiedener Wörter interessant, in denen das Trigramm auftritt. Denn: diese Anzahl sagt ihm, wieviele Wörter ihm im nächsten Schritt als weitere Terme präsentiert werden, wenn er das Trigramm als domänenspezifisch bestätigt (es werden ja alle Wörter herausgesucht, die das Trigramm enthalten, siehe Abschnitt 3.3.1).

Konzeptuell aber gibt es so gut wie keinen Unterschied zwischen der Differenzanalyse für Wörter und der für Buchstabentrigramme.

3.2.4 Eigennamen

Je nach Textsorte kann es vorkommen, daß ein beträchtlicher Teil der mittels Differenzanalyse ermittelten Terminologie aus Eigennamen besteht. Diese als solche zu erkennen, ist aus mehreren Gründen interessant:

- In wissenschaftlichen Artikeln treten Eigennamen bevorzugt in Zitaten (als Autoren anderer Artikel) auf. Sie weisen zwar eine hohe Fachspezifik im Sinne eines hohen P-Wertes auf, gehören aber sicherlich nicht zum „Begriffs- und Benennungssystem“ der Domäne. Insofern will man sie vermutlich als nicht relevant verwerfen.

- In Konzernberichten oder sonstigen Texten aus dem Bereich der Wirtschaft treten Eigennamen oft als Produkt- oder Firmennamen in Erscheinung. Diese können einerseits wirklich interessante Terminologie darstellen, andererseits sind sie oft auch ganz normale Wörter: hier ist es z.B. für ein maschinelles Übersetzungssystem entscheidend wichtig zu wissen, daß sie in dieser Domäne nicht übersetzt werden dürfen.

  Ein Beispiel: Der Name der Softwarefirma „Sun“ sollte bei der Übersetzung eines englischen Textes *nicht* mit „Sonne“ ins Deutsche übertragen werden, sondern unverändert bleiben.

Es gibt sicher noch mehr Gründe, die dafür sprechen, Eigennamen als solche zu identifizieren und getrennt von der übrigen Terminologie anzuzeigen. Daher will ich dies versuchen — allerdings ohne großen Aufwand zu treiben.

Die von mir implementierte Eigennamenerkennung stützt sich nur auf die POS-Tags des TnT: alle Wörter, die bei *jedem* ihrer Auftreten das Tag „NE“

(für „named entity“) erhalten haben, werden als Eigennamen identifiziert und getrennt angezeigt.

So werden zwar in den meisten Fällen *nicht alle* Eigennamen erkannt. Diejenigen Wörter, die als Eigennamen eingestuft werden, sind aber mit hoher Sicherheit auch solche.

Die getrennte Anzeige hilft jedenfalls bei der schnelleren Bearbeitung der Liste der Einwortterme: Eigennamen sind — wie oben erwähnt — oft keine Terminologie, so daß man in vielen Domänen den Abschnitt mit Eigennamen einfach überspringen kann.

3.3 Schritt 2: verschiedene Pendel

Hat man die Differenzanalyse in Schritt 1 durchgeführt und hat der Anwender die Fachtermini und domänenspezifischen Buchstabentrigramme seiner Wahl markiert, so kann man diese verwenden, um nach weiteren Fachtermini zu suchen. Im Folgenden soll diese anfängliche Menge von Termen und Trigrammen mit *Ausgangsmenge* (oder englisch *seed list*) bezeichnet werden.

Das, was nun folgt, möchte ich „Pendeln“ nennen, denn es soll so etwas stattfinden wie ein Sich-Aufschaukeln: Im ersten Pendelschritt wird die Ausgangsmenge erweitert; für den zweiten werden auch die im ersten Pendelschritt neu gefundenen Termini miteinbezogen usw. Das Ganze kann man (wenn der Benutzer genug Geduld hat) solange durchführen, bis keine weiteren Fachbegriffe mehr gefunden werden.

Dabei wird in drei verschiedene Richtungen gesucht: Ein morphologisches Pendel sucht nach Wörtern, die domänenspezifische Trigramme enthalten. Ein syntaktisches Pendel sucht nach Phrasen bzw. bestimmten POS-Mustern, die möglichst schon gefundene Terme enthalten sollen. Und ein semantisches Pendel sucht nach weiteren Einwortthermen, die denen der Ausgangsmenge semantisch ähnlich sind. Zur Veranschaulichung des Ablaufs betrachte man sich nocheinmal Abb. 3.1.

Die einzelnen Pendel sollen im Folgenden näher beschrieben werden.

3.3.1 Morhologisches Pendel

Das morphologische Pendel ist sehr einfach: es tut nichts weiter, als Wörter aus dem Fachtext zu extrahieren, die domänenspezifische Buchstabentrigramme der seed list enthalten.

Hierzu ein Beispiel: In einem medizinischen Text ist das Trigramm „cyt“ als domänenspezifisch erkannt worden. Das morphologische Pendel findet alle Wörter, die „cyt“ enthalten:

> *lymphocyte, antithymocyte, thrombocytopenia, cytokine, b-lymphocyte, cytomegalovirus*

Im Gegensatz zu den anderen Pendeln gibt es hier keinerlei Filter, insbesondere keine Mindestfrequenz. Man stellt nämlich sehr schnell fest, daß ein großer Teil der so gefundenen Wörter nur einmal im Text vorkommt: von den hier aufgeführten tritt nur „lymphocyte“ mehrfach auf.

Obwohl sie so selten und damit wenig repräsentativ für den Inhalt des Textes sind, handelt es sich bei diesen Wörtern sehr oft um Fachtermini. Bei der Differenzanalyse könnte man diese Terme nie finden, da man dort ohne Mindestfrequenz nicht auskommt. Das morphologische Pendel ist also ein wichtiger Schritt bei der Erhöhung des Recalls der Extraktion, ohne daß man dabei auf eine hohe Precision verzichten muß.

Allerdings funktioniert das morphologische Pendel nur bei bestimmten Textsorten richtig gut: lediglich in Fachgebieten, in denen verstärkt lateinische oder griechische Morphologie benutzt wird, ist die Precision wirklich hoch, bzw. nur dort findet man auch eine große Anzahl domänenspezifischer Trigramme.

Läßt es sich automatisch, d.h. ohne Eingreifen des Anwenders voraussagen, ob ein Text (dessen Fachdomäne dem System ja zunächst unbekannt ist) für das morphologische Pendel geeignet ist oder nicht? Die Vermutung liegt nahe, daß bei der Durchführung der Differenzanalyse für die Trigramme durchschnittlich höhere Werte der statistischen Prüfgröße zu erwarten sind, wenn es sich um eine der „guten“ Domänen wie Medizin handelt.

Das heißt: das System sollte sehr wohl in der Lage sein, zu beurteilen, ob sich die Anwendung des morphologischen Pendels lohnt. Und falls es sich

nicht lohnt, sollten dem Anwender erst gar keine Trigramme zur Durchsicht vorgelegt werden.

Diese „Domänenerkennung“ habe ich zunächst nicht implementiert, die Anpassung solcher Parameter soll jedoch in Kapitel 6 untersucht werden.

3.3.2 Syntaktisches Pendel

Das syntaktische Pendel dient dem Auffinden von Mehwortbegriffen. Es ist ein sehr wichtiger Bestandteil des Systems, da sehr viele Fachtermini des Englischen aus mehr als einem Wort bestehen.

Bei der Umsetzung griff ich die Ideen von [Daille 1994], [Frantzi 1996] und [Justeson 1995] auf (vgl. Abschnitt 2.2.2): von Daille und Justeson habe ich die Überzeugung übernommen, daß die pure Frequenz einer Phrase das beste Auswahlkriterium ist, von Frantzi das Maß *C-Value*. Insbesondere wird bei meiner Implementierung das Problem der *formalen Variation* (bewußt) ignoriert.

Die Extraktion von Phrasen geht also folgendermaßen vor sich:

Zunächst wird nach POS-Mustern gesucht, die der Anwender selbst bestimmen kann. Allerdings werden ihm dabei Vorschläge gemacht, die sowohl (nach [Arppe 1995]) gute Ergebnisse versprechen, als auch die Anwendung des *C-Value* erleichtern.

Zur Erinnerung: der *C-Value* einer Phrase sinkt, wenn diese fast ausschließlich innerhalb anderer, längerer Phrasen auftritt. Daher ist es sinnvoll, zuerst die langen Phrasen zu suchen, um für die kürzeren den *C-Value* korrekt berechnen zu können.

Die vorgeschlagenen POS-Muster sind im ersten Pendelschritt:

N N N A N N A A N

wobei A wieder für Adjektiv, N für Nomen steht. Im zweiten Pendelschritt wird dann nach den Mustern

A N und N N

gesucht.

Bisher kann man noch nicht von „Pendel“ sprechen, da die Elemente der seed list nicht in die Auswahl der Phrasen miteinfließen. Die vom Benutzer

bereits bestätigten Terme stellen aber eine wichtige Information dar, da ihre Fachspezifik (*termhood*) schon feststeht. Daher wird der *C-Value* einer Phrase mit folgendem Wert multipliziert:

$$phrase\_termhood = 1 + \frac{\#\text{Terme}}{\#\text{Wörter}} \tag{3.15}$$

Der *C-Value* einer Phrase aus drei Wörtern, welche zwei bereits bekannte Terme enthält, wird beispielsweise mit dem Wert $1 + \frac{2}{3} = 1.67$ multipliziert. Allgemein liegt der Wert von *phrase_termhood* zwischen 1 (wenn die Phrase keine bekannten Terme enthält) und 2 (wenn sie nur aus solchen besteht). Somit werden Phrasen, die bekannte Terme enthalten, als besser bewertet.

Zusätzlich gibt es noch einen Filter, der Phrasen mit „schlechten“ Adjektiven aussortiert: ein Adjektiv gilt als schlecht, wenn es nur eine textstrukturierende Funktion bzw. eine sehr allgemeine, unspezifische Bedeutung hat. Beispiele für schlechte Adjektive sind: *following, other, corresponding, entire, whole, ...*

Die gefundenen Phrasen werden — absteigend nach *C-Value* geordnet — zusammen mit diesem, ihrer Frequenz und zwei Beispielsätzen ausgegeben.

3.3.3 Semantisches Pendel

Das semantische Pendel sucht Wörter, welche zu Elementen der Ausgangsmenge in statistischer paradigmatischer Relation stehen.

Aufzählungen

So wie in Abschnitt 1.4.3 beschrieben, wird dazu zunächst mittels POS-Mustern nach Aufzählungen gesucht. Die verwendeten Muster sowie zugehörige Beispiele sind in Tabelle 3.1 dargestellt.

Im Folgenden werden nur noch Aufzählungen betrachtet, bei denen alle — bis auf eines — der beteiligten Nomina bereits Termstatus haben. Das fehlende Nomen erhält dann ebenfalls Termstatus, da der starke Verdacht besteht, daß es zu den anderen Nomina der Aufzählung in paradigmatischer Beziehung steht, diesen also semantisch ähnlich ist. Und Wörter, die eine ähnliche Bedeutung haben wie bereits gefundenen Fachtermini, sind vermutlich auch Terminologie.

POS-Muster	Beispiel-Aufzählung
N C N	globulin and azathioprine
N , N C N	globulin, azathioprine and thalidomide
N C A N	fever and other symptoms
N , N C A N	anorexia, fever and other symptoms

Tabelle 3.1: POS-Muster für die Extraktion von Aufzählungen mit zugehörigen Beispielen; C steht dabei für Konjunktionen wie „and“.

Kollokationen

Eine weitere Art, semantisch ähnliche Wörter zu Elementen der Ausgangsmenge zu finden, ist das Rechnen mit Kollokationsvektoren. Verwendet wird dabei der in Tabelle 1.1 beschriebene Algorithmus, der zu den bereits gefundenen Termen Wörter mit ähnlichen Satzkollokationen sucht.

Dazu werden zunächst alle Satzkollokationen berechnet. Hierzu verwendete ich ein an der Universität Leipzig entwickeltes Programm, welches das oben bereits erwähnte Poisson-Maß zur Berechnung von Kollokationen benutzt (vgl. wieder [Quasthoff 2002]):

Zu zwei Wörtern A und B und einem Text mit n Sätzen lautet das statistische Experiment hier: für einen Satz prüfe man, ob die Wörter A und B gemeinsam in diesem auftreten (positives Ergebnis) oder nicht (negativ). Das macht man für alle n Sätze und interessiert sich für die Wahrscheinlichkeit k positiver Ergebnisse.

Die Nullhypothese ist hier die statistische Unabhängigkeit der Wörter A und B, d.h. man geht davon aus, daß die Wahrscheinlichkeit $p(A, B)$, beide Wörter gemeinsam in einem Satz anzutreffen, gleich $p(A)p(B)$ ist.

Damit ergibt sich für n Sätze ein Erwartungswert von $\lambda = np(A)p(B)$ gemeinsamen Auftreten. Die Wahrscheinlichkeiten p(A) und p(B) werden dabei wieder als relative Häufigkeiten aus dem Text geschätzt.

Nun kann man also mittels Formel 3.13 (oder der exakteren Formel (3.12), je nach Größenordnung von λ) das „Erstaunen“ darüber messen, A und B k-mal gemeinsam in einem Satz gesehen zu haben, wenn man λ als Erwartungswert (und die zugehörige Poisson-Verteilung) voraussetzt.

Die Satzkollokationen eines Wortes w sind also diejenigen Wörter w_i, für die $sig(w, w_i)$ besonders groß ist. Diese werden als Vektor abgespeichert, indem zu jedem w_i an die entsprechende Stelle im Vektor der Wert $sig(w, w_i)$ geschrieben wird. Der ganze Vektor wird dann normiert, d.h. alle Einträge werden durch den Betrag des Vektors dividiert.

Um Rechenzeit zu sparen, werden nicht die Ähnlichkeiten *aller* Wörter des Textes zu Elementen der seed list ausgerechnet, sondern eine Menge von Kandidaten ausgewählt. Und zwar enthält diese Menge alle starken Kollokationen von Elementen der Ausgangsmenge.

Nun wird für jeden Kandidaten seine Ähnlichkeit zu allen bereits gefundenen Termini berechnet. Ist $\vec{t}$ der Vektor der Satzkollokationen eines Terms t, $\vec{k}$ der eines Kandidaten k, so berechnet sich die Ähnlichkeit zwischen beiden Wörtern über das Cosinus-Maß (bzw. das Skalarprodukt, was hier dasselbe ergibt, da die Vektoren normiert sind):

$$sim(t, k) = \sum_{i=1}^{n} t_i k_i \tag{3.16}$$

n bezeichnet hier die Anzahl aller Wörter, die im Text vorkommen.

Das heißt: die Ähnlichkeit ist groß, wenn die Vektoren möglichst viele möglichst große Werte gemeinsam haben; oder anders gesagt: wenn die Wörter t und k viele starke Kollokationen teilen. Diese Gemeinsamkeit bezüglich des Kontexts (also der Kollokationen) hatte ich ja im Abschnitt 1.4.3 als statistische paradigmatische Beziehung definiert.

Extrahiert werden nun alle Wörter, die zu *irgendeinem* der Elemente der Ausgangsmenge eine Ähnlichkeit aufweisen, die einen Schwellenwert S überschreitet.

Einen guten Wert für S habe ich dabei zunächst durch Ausprobieren gefunden. Der so festgelegte Wert liefert zwar ordentliche Ergebnisse, es ist jedoch nicht klar, ob seine Größenordnung vielleicht vom Umfang oder der Art des Eingabetextes abhängt und somit dynamisch angepaßt werden muß. Dies soll in Kapitel 6 untersucht werden.

Zunächst aber möchte ich im folgenden Kapitel die hier beschriebene Implementierung unter Berücksichtigung verschiedener Aspekte evaluieren.

Hieraus sich eventuell ergebende weitere Defizite des Systems sollen dann ebenfalls in Kapitel 6 behandelt werden.

Kapitel 4

Erste Ergebnisse

Gegenstand dieses Kapitels soll die Evaluierung des von mir implementierten Systems sein. Diese werde ich mit den im Information Retrieval üblichen Verfahren und Maßen durchführen: es werden Precision, Recall und F-Wert berechnet, um die Leistungsfähigkeit des Systems zu messen. Zunächst also eine genaue Definition der drei Maße: Sei G die Menge aller vom System gefundenen Wörter und R die Menge aller relevanten (d.h. tatsächlich vorhandenen) Fachtermini.

- Die Precision eines Suchergebnisses berechnet sich dann als

$$P = \frac{|G \cap R|}{|G|} \tag{4.1}$$

  Das heißt die Precision gibt an, wieviel Prozent der gefundenen Wörter wirklich Terme sind

- Der Recall ist gegeben durch:

$$P = \frac{|G \cap R|}{|R|} \tag{4.2}$$

  Er gibt an, wieviel Prozent aller vorhandenen Terme vom System gefunden wurden.

- Schließlich dient der F-Wert dazu, Precision und Recall zu einem Maß zusammenzufassen. Er berechnet sich als das harmonische Mittel aus P und R:

$$F = \frac{2PR}{P + R} \tag{4.3}$$

Der F-Wert ist ein Maß für die Gesamtgüte des Systems, da er nur dann groß wird, wenn sowohl P als auch R einen hohen Wert annehmen.

In [Riloff 1994] findet sich eine Variante des F-Wertes, die es zuläßt, Precision und Recall unterschiedlich zu gewichten:

$$F(\beta) = \frac{(1+\beta^2)PR}{\beta^2 P + R} \tag{4.4}$$

Für $\beta = 1$ ergibt sich der ursprüngliche F-Wert, bei dem Precison und Recall gleiches Gewicht haben. Für $\beta = 0,5$ ist P doppelt so wichtig wie R, für $\beta = 2$ umgekehrt. Da ich mehr Wert auf die Precision des Systems lege, werde ich im folgenden das Maß $F(0,5)$ verwenden.

Als nächstes stellt sich die Frage, wie festgelegt wird, welche Wörter und Phrasen die Terminologie eines Textes darstellen. Wie in Abschnitt 1.2 erklärt wurde, ist diese Abgrenzung eigentlich vom Anwender abhängig und daher subjektiv. Im Folgenden sollen aber zwei Szenarien entworfen werden, die es erlauben, die Entscheidung ein Stück weit zu objektivieren.

4.1 Testszenarien

Zwei wichtige Anwendungen der Terminologie-Extraktion sollen als Testszenarien zum Einsatz kommen: die maschinelle Übersetzung und die Erstellung von Fachwörterbüchern. Innerhalb dieser Anwendungen ergibt sich die Frage danach, was Terminologie ist, wie folgt:

- Maschinelle Übersetzung: Der Eingabetext wird von dem Übersetzungssystem LOGOS übersetzt. Alle Wörter oder Phrasen, die dabei nicht oder falsch übersetzt werden (weil sie im Lexikon von LOGOS nicht oder nur mit einer für die Domäne nicht passenden Übersetzung enthalten sind), werden als Terminologie angesehen.

- Erstellung eines Fachwörterbuchs: Ein Experte liest den gesamten Eingabetext und streicht alle Wörter oder Phrasen an, die er in ein Fachwörterbuch der betreffenden Domäne aufnehmen würde. Dabei empfiehlt es sich, ein bereits vorhandenes Fachwörterbuch zum Abgleich heranzuziehen.

Beim Übersetzen mit LOGOS kann man dabei so etwas wie eine *Baseline* annehmen, gegeben durch ein (fiktives) System, das einfach nur alle Wörter des Fachtextes im LOGOS-Lexikon nachschlägt und alle extrahiert, die nicht enthalten sind. Man muß sich aber im Klaren darüber sein, daß ein solches System keinerlei Mehrwortbegriffe finden kann.

Ich habe beide Experimente jeweils mit zwei englischen Texten durchgeführt:

- mit einem medizinischen Text, welcher die verschiedenen Komplikationen und Nachwirkungen bei Knochenmarkstransplantationen beschreibt und
- mit einem linguistischen Text, der eine Einführung in die Begriffe der Morphologie gibt.

Als Expertin zog ich im ersteren Fall eine Studentin der Medizin heran. Begleitend benutzten wir das „Taschenwörterbuch der Medizin“ (P. und C. Reuter, Thieme Verlag) für Klärung von Übersetzungsfragen und das Klinische Wörterbuch „Pschyrembel“ (de Gruyter) als Fachwörterbuch.

Für den linguistischen Text betrachtete ich mich selbst als Experten und verwendete das „Lexikon der Sprachwissenschaft“ ([Bußmann 2002]) als Fachwörterbuch.

Ein Experiment mit einem Text aus dem Bereich der Informatik zeigte schnell, daß das verwendete Referenzkorpus für Texte dieser Domäne tatsächlich (s. Abschnitt 3.2.2) ungeeignet ist: die Differenzanalyse versagte weitgehend.

Nachdem die Vorarbeit erledigt war — nämlich alle Fachtermini aus beiden Texten manuell zu extrahieren — konnte nun die eigentliche Evaluierung beginnen. Diese soll zunächst für das Gesamtsystem und dann für dessen Komponenten durchgeführt werden.

4.2 Gesamtsystem

Meine Implementierung der Terminologie-Extraktion beruht — wie in Kapitel 3 beschrieben — auf Interaktion mit dem Benutzer. Dazu gehört auch,

daß dieser verschiedene Parameter frei variieren kann, wie z.B. die Mindestfrequenz für Terme bei der Differenzanalyse, Schwellenwerte für die statistische Prüfgröße usw. Auch die Auswahl der richtigen Termini nach jedem Extraktionsschritt verbessert das Verhalten des Systems im nächsten Schritt.

Auf diese Interaktion soll aber für die Evaluierung verzichtet werden (da sie schwer in Zahlen zu fassen ist). Stattdessen habe ich für alle verstellbaren Parameter zunächst diejenigen Werte angenommen, die dem Anwender als Defaults vorgeschlagen werden. Diese hatte ich vorher nach einigem „Herumprobieren“ für gut befunden.

Bei der Analyse der Systemkomponenten (siehe Abschnitt 4.3) möchte ich dann das „Herumprobieren“ systematisieren, d.h. genau untersuchen, wie sich Precision und Recall der Komponenten verändern, wenn man die Parameter systematisch variiert.

Zunächst aber führte ich die Evaluierung mit folgenden Parameterwerten durch:

- statistische Prüfgröße: Häufigkeitsquotient; Schwellenwert für Wörter: 1000 (d.h. es werden alle Wörter extrahiert, deren relative Häufigkeit im Fachtext 1000-mal größer als im Referenzkorpus ist), für Trigramme: 25
- Mindestfrequenz für Wörter bei der Differenzanalyse: 2
- Mindestanzahl Wörter, in denen ein Trigramm vorkommen muß (Differenzanalyse für Trigramme): 3
- Schwellenwert für C-Value bei der Phrasenextraktion: 2
- Mindestähnlichkeit für Terme, die mittels semantischem Pendel gefunden werden: 0.35

Es zeigte sich auch, daß beim semantischen Pendel ein kleiner zusätzlicher Filter verwendet werden muß, um die Qualität der Ergebnisse zu sichern: ein Mindestwert der statistischen Prüfgröße, der allerdings wesentlich kleiner als bei der Differenzanalyse gewählt werden konnte (ich setzte ihn auf 50, also ein zwanzigstel des Schwellenwertes der Differenzanalyse).

Akzeptiert wurden dann alle vom System gefundenen Wörter (außer die als Eigennnamen gekennzeichneten); diese flossen dann auch jeweils in die nächsten Schritte mit ein. Nur bei den Buchstabentrigrammen griff ich ein, indem ich die besten auswählte.

Zunächst noch ein paar Einzelheiten zu den Texten: Der medizinische Text besteht aus 5766 Wortformen (davon 1162 verschiedene); vom „Fachmann“ wurden 422 verschiedene Fachtermini (davon 133 Phrasen) identifiziert, in LOGOS fehlten 280 (davon 77 Phrasen). Mein System fand 295 Wörter bzw. Phrasen.

Der linguistische Text umfaßt 4809 Wortformen (davon 952 verschiedene); nach Expertenmeinung gab es in diesem Text 133 Fachtermini (40 Phrasen), LOGOS irrte sich bei der Übersetzung von 73 Wörtern (18 Phrasen). 130 Wörter bzw. Phrasen wurden von meinem System gefunden.

Es fällt sofort auf, daß der medizinische Text wesentlich mehr Fachbegriffe enthält als der linguistische, obwohl beide Texte ungefähr gleich lang sind. Das liegt daran, daß im medizinischen Text sehr viele Bereiche der Medizin gestreift werden: die Folgen der Knochenmarktransplantation erstrecken sich auf den ganzen Körper (Lunge, Herz, Zähne, Augen, Ohren und Knochen, um nur einige Beispiele zu nennen). Aus allen zugehörigen Spezialgebieten der Medizin fließt so Terminologie in den Text ein. Sehr viele der Fachtermini (41%) kommen daher nur einmal vor, da sie als Hintergrundwissen vorausgesetzt werden.

Der linguistische Text setzt demgegenüber keinerlei linguistische Vorkenntnisse voraus. Alle eingeführten Begriffe werden zunächst definiert und später meist wiederverwendet (83% der Fachtermini kommen mehr als einmal im Text vor). Dieser grundsätzliche Unterschied der beiden Texte wird uns später noch ausführlicher beschäftigen.

Precision und Recall für beide Texte und Testszenarien sind in Tabelle 4.1 dargestellt.

Seltene Fachbegriffe (insbesondere solche, die nur einmal im Text auftreten) bereiten Probleme bei der Differenzanalyse: sie können aufgrund der Mindestfrequenz nicht gefunden werden. Es verwundert daher wenig, daß der medizinische Text einen schlechteren Recall liefert als der linguistische.

Die Ergebnisse beim Abgleich mit LOGOS sind insgesamt enttäuschend:

	Fachmann	LOGOS	LOGOS-Baseline
Morphologie	$P = 52\%$ $R = 51\%$	$P = 27\%$ $R = 48\%$	$P = 21\%$ $R = 41\%$
Medizin	$P = 66\%$ $R = 46\%$	$P = 44\%$ $R = 47\%$	$P = 62\%$ $R = 58\%$

Tabelle 4.1: Precision und Recall des Gesamtsystems für verschiedene Testszenarien

beim medizinischen Text liegen sie sogar unter der Baseline. Das mag man einerseits mit dem Problem der seltenen Terme erklären (das die Baseline nicht hat, da sie nur Wörter im Lexikon nachschlägt). Andererseits sind die Ergebnisse (auch die der Baseline) insgesamt schlecht, so daß man vermuten könnte, daß diese Art der Terminologie-Extraktion für das Auffüllen von Lexika zur maschinellen Übersetzung ungeeignet ist.

Wie sich die Ergebnisse insgesamt verbessern lassen, soll nun untersucht werden, indem die einzelnen Komponenten des Systems analysiert und die jeweils besten Einstellungen der Parameter ermittelt werden. Dabei wird es insbesondere interessant sein zu erfahren, ob die optimalen Einstellungen für beide Texte gleich sind. Dann nämlich bestünde Hoffnung darauf, global (d.h. für alle Texte geltende) optimale Parameterwerte finden zu können.

4.3 Komponentenanalyse

4.3.1 Differenzanalyse

Vergleich der Prüfgrößen

Als erstes möchte ich die verschiedenen statistischen Prüfgrößen miteinander vergleichen. Dabei werde ich mich im Folgenden auf die Expertenurteile als Vergleichsgröße beschränken. Die Experimente mit LOGOS lieferten jeweils qualitativ ähnliche Ergebnisse.

Ich führte die Differenzanalyse zunächst mit dem Häufigkeitsquotienten (HQ) mit den oben beschriebenen Parametern durch. Dann stellte ich für

die Likelihood ratio (LR) und das Poisson-Maß den Schwellenwert so ein, daß die Ergebnismenge genauso groß wurde wie beim Häufigkeitsquotienten. Die Mindestfrequenz beließ ich in allen Fällen bei 2.

Die Precision- und Recallwerte, die sich dabei ergaben, sind in Tabelle 4.2 dargestellt. Der Recall bezieht sich dabei nur auf die vom Fachmann identifizierten *Einwortterme*, nicht auf Phrasen.

	HQ	LR	Poisson
Morphologie	$P = 65\%$ $R = 32\%$	$P = 81\%$ $R = 41\%$	$P = 80\%$ $R = 40\%$
Medizin	$P = 84\%$ $R = 25\%$	$P = 66\%$ $R = 19\%$	$P = 67\%$ $R = 19\%$

Tabelle 4.2: Precision und Recall bei der Differenzanalyse

Das erste, was auffällt, sind die sehr ähnlichen Ergebnisse der LR und des Poisson-Maßes. In der Tat sind die Ergebnismengen beider Maße identisch; nur ließ sich der Schwellenwert nicht so wählen, daß genau gleich viele Terme gefunden wurden. Beim medizinischen Text z.B. umfaßte die Ergebnismenge der LR 85 Wörter, die des Poisson-Maßes 81. Alle diese 81 Wörter waren — und zwar in der gleichen Reihenfolge, wenn man sie nach Prüfgrößenwert ordnet — in der Menge der 85 LR-Ergebnisse enthalten.

Es bleibt also nur noch der Vergleich zwischen HQ und LR. Hier sind die Ergebnisse sehr unterschiedlich. Im medizinischen Beispiel fanden beide Maße 85 Terme; davon stimmten nur 40 (47%) überein. Die Erklärung hierfür liegt in der oben aufgestellten These, der HQ bevorzuge (bzw. überbewerte) seltene Terme. Der Mittelwert der Frequenzen aller vom HQ gewählten Terme liegt bei 5,8, die durchschnittliche Frequenz der LR-Terme bei 13,9. Auch ohne statistische Signifikanztests kann man sagen, daß hier eine erhebliche Abweichung vorliegt.

Interessant ist aber, daß der HQ bezüglich Precision und Recall beim medizinischen Text wesentlich besser abschneidet als die LR; beim linguistischen Text ist es genau umgekehrt. Auch dieses Phänomen kann mit der genannten These erklärt werden: wie schon festgestellt kommen 41% der vom Fachmann im medizinischen Text identifizierten Fachtermini nur einmal vor;

beim linguistischen Text sind es nur 17%. Die vermeintliche Schwäche des HQ (nämlich seltene Ergebnisse überzubewerten) ist also eine Stärke, wenn Texte wie der medizinische untersucht werden, die viel Hintergrundwissen voraussetzen und damit viele Terme enthalten, die nur einmal auftreten.

Welches Maß bessere Ergebnisse liefert, läßt sich also nicht global feststellen, sondern ist abhängig von der Textsorte. Einführende Texte, die kein Vorwissen voraussetzen und somit die eingeführten Begriffe wiederholen, lassen sich besser mit der LR bearbeiten, Zusammenfassungen mit fachspezifischem Hintergrundvokabular besser mit dem HQ. Die Frage, ob automatisch entschieden werden kann, welches Maß zu verwenden ist oder ob der Benutzer das Maß selbst wählen muß, soll in Kapitel 6 geklärt werden.

Variation der Parameter

In diesem Abschnitt möchte ich der Frage nachgehen, ob die anfangs gewählten Default-Werte für den Schwellenwert der Prüfgröße und die Mindestfrequenz optimal waren bzw. welches die optimalen Werte sind. Dabei werde ich von nun an nur noch die LR als Prüfgröße untersuchen. Denn trotz der verschiedenen Ergebnisse waren die sich ergebenden Kurven für den HQ qualitativ gleichwertig.

Zunächst muß geklärt werden, ob — wie in Abschnitt 1.3 behauptet — die Verteilung von Fachtermini in Texten wirklich signifikant anders ist als die von „Nichttermen“. Abbildung 4.1 und 4.2 zeigen die Verteilung der LR bei Termen und Nichttermen für den medizinischen Text.

Man sieht, daß die Verteilung wirklich sehr unterschiedlich ist: bei den Nichttermen gibt es nur wenige, deren LR-Wert über 20 liegt, bei der Mehrzahl der Terme ist dies hingegen der Fall. Die Verteilung bei den Nichttermen folgt grob dem Zipfschen Gesetz, während die Kurve bei den Termen an eine Binomialverteilung erinnert.

Dies macht Mut: wählt man den Schwellenwert der Prüfgröße größer als beispielsweise 30, so kann man mit einer recht guten Precision rechnen. Andererseits sieht man auch, daß eine große Anzahl von Termen einen LR-Wert unter 30 hat, sodaß man nicht mit erschöpfendem Recall rechnen kann. Das heißt, man hat den üblichen Trade-off zwischen Precision und Recall:

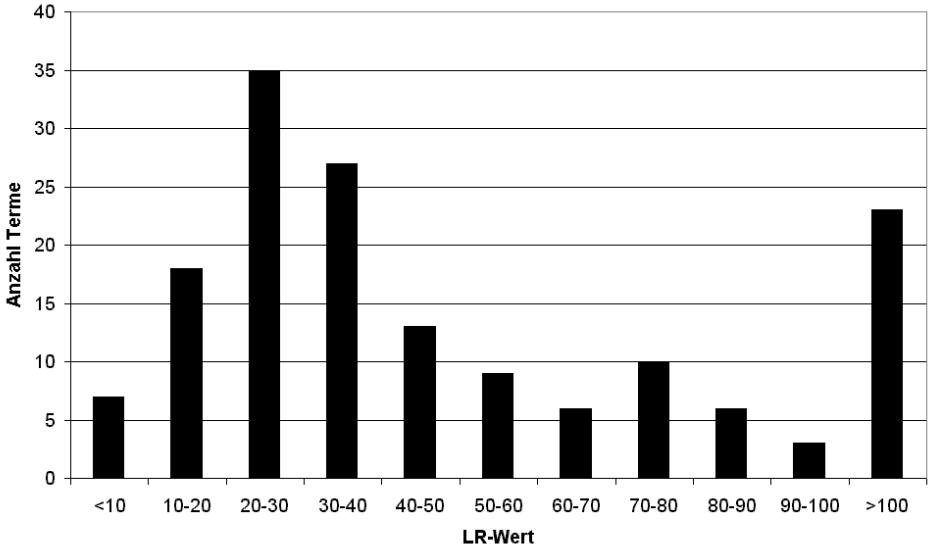

Abbildung 4.1: Verteilung der LR bei Termen

ein hoher Schwellenwert hat große Precision und kleinen Recall zur Folge, ein niedriger umgekehrt.

Abbildung 4.5 zeigt die Entwicklung von Precision, Recall und $F(0,5)$ bei Variation des Schwellenwertes für die LR in Schritten von jeweils 2,5 (ebenfalls für den medizinischen Text). Die Mindestfrequenz wurde hier bei 2 festgehalten.

Erfreulicherweise hat der F-Wert ein klares Optimum (bei LR = 22,5); dort liegt die Precision bei 60%, der Recall bei 44%. Bei meinen ersten Experimenten hatte ich den Schwellenwert bei 53 gewählt, also offensichtlich zu hoch.

Bezieht man die Mindestfrequenz mit in die Betrachtung ein, so ergibt sich ein dreidimensionales Problem: Abb. 4.3 zeigt den F-Wert als Funktion des Schwellenwertes für die LR *und* der Mindestfrequenz für Terme. Dies ist das Bild für den medizinischen Text. Man sieht, daß das Optimum des F-Wertes bei einer Mindestfrequenz von 1 erreicht wird. Dies ist aber nur für den medizinischen Text so: Abb. 4.4 zeigt dasselbe für den linguistischen Text. Hier liegt das Optimum bei einer Mindestfrequenz von 2. Wieder stellt sich heraus, daß der medizinische Text sehr viele seltene Be-

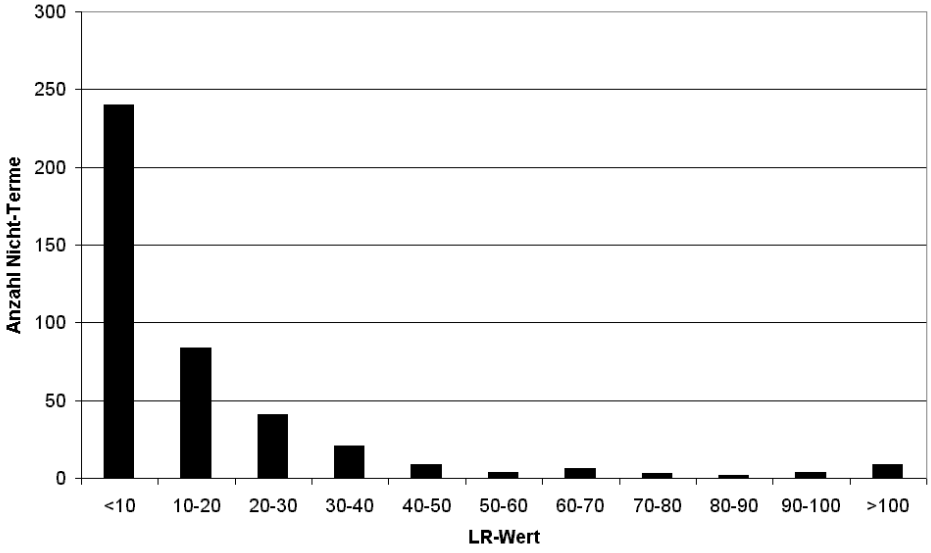

Abbildung 4.2: Verteilung der LR bei Nichttermen

griffe enthält, sodaß hier eine niedrigere Mindestfrequenz angebracht ist als beim linguistischen Text.

Tabelle 4.3 zeigt schließlich noch für beide Texte und beide Prüfgrößen HQ und LR diejenigen Paare aus Mindestfrequenz (min_freq) und Schwellenwert (min_P), für die der F-Wert sein Optimum annimmt (bzw. die drei besten Einstellungen). Interessant ist, daß die LR als Prüfgröße zwar für den linguistischen Text weiterhin bessere Ergebnisse liefert als der HQ, daß der Unterschied jedoch kleiner geworden ist. Der HQ schneidet hingegen für den medizinischen Text weiterhin wesentlich besser ab als die LR.

Leider gibt es offensichtlich weder bei der Mindestfrequenz noch beim Schwellenwert der Prüfgröße eine Einstellung, die für *beide* Texte optimal ist. Die Unterschiede beider Texte spiegeln sich wie erwartet wider: der medizinische Text braucht bei beiden Parametern niedrigere Werte als der linguistische.

Außerdem fällt auf, daß der HQ insgesamt mit größeren Mindestfrequenzen verwendet werden muß als die LR: auch dies ist klar, wenn man sich daran erinnert, daß der HQ seltene Ereignisse überbewertet; das muß durch eine entsprechend höhere Mindestfrequenz ausgeglichen werden.

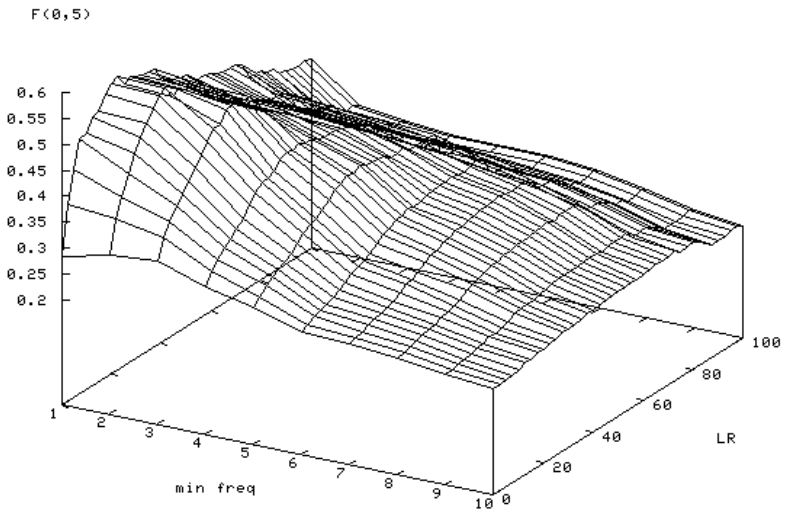

Abbildung 4.3: $F(0,5)$ als Funktion der LR und der Mindestfrequenz (medizinischer Text)

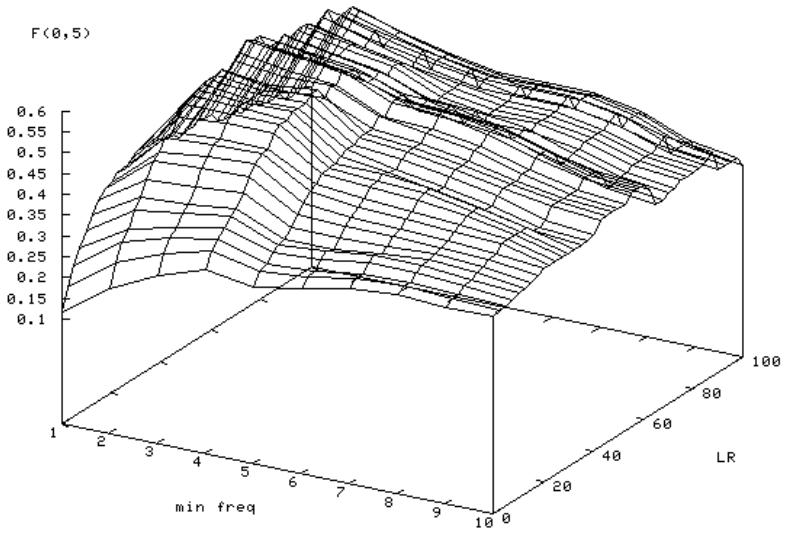

Abbildung 4.4: $F(0,5)$ als Funktion von LR und Mindestfrequenz (linguistischer Text)

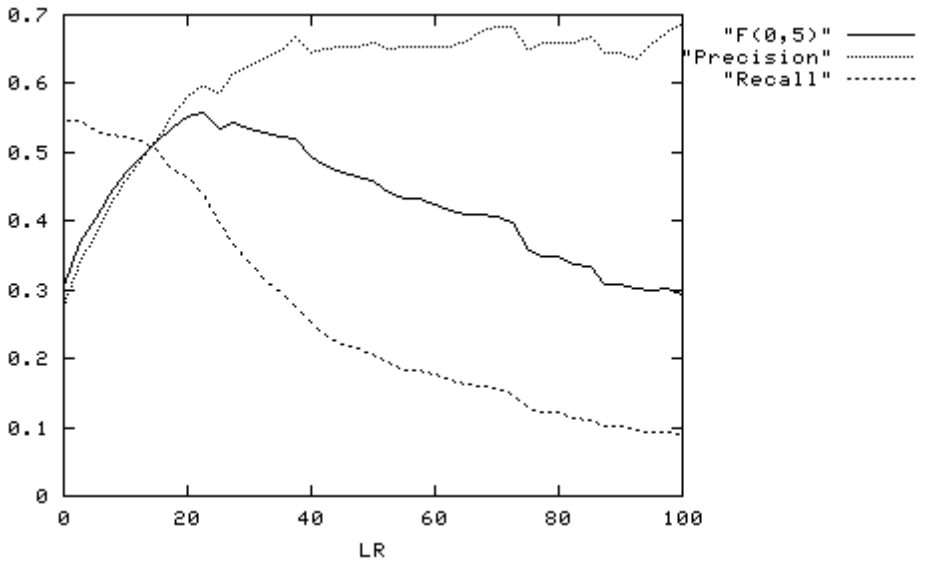

Abbildung 4.5: Precision, Recall und $F(0,5)$ bei Variation der LR

	HQ		LR	
	(min_freq; min_P)	F-Wert	(min_freq; min_P)	F-Wert
Morphologie	(4; 100)	0,58	(2; 55)	0,59
	(3; 100)	0,56	(2; 57,5)	0,58
	(3; 400)	0,54	(2; 60)	0,58
Medizin	(2; 200)	0,67	(1; 22,5)	0,56
	(2; 300)	0,66	(2; 22,5)	0,56
	(2; 250)	0,66	(1; 20)	0,56

Tabelle 4.3: Paare (min_freq; min_P), für die $F(0,5)$ optimal ist

4.3.2 Morphologisches Pendel

Bei der Analyse des morphologischen Pendels ging ich zunächst von den in Abschnitt 4.2 beschriebenen Parametern aus: ich verwendete den HQ als Prüfgröße mit einem Schwellenwert von 25. Ein weiterer Parameter der Differenzanalyse für Buchstabentrigramme ist die Mindestanzahl M von Wörtern, in denen ein Trigramm vorkommen muß (nicht die Frequenz des Trigramms an sich). M hatte ich auf 3 gesetzt. Auch hier beschränkte ich mich

auf den Vergleich mit den vom Experten gefundenen Termen.

Die Default-Einstellungen lieferten für den medizinischen Text 84,2% Precision, für den linguistischen 68,4%. Der Recall war in beiden Fällen gering, was aber auch daran liegt, daß das morphologische Pendel nur Terme ausgibt, die nicht bereits in der Differenzanalyse gefunden wurden. Es ist aber zu beachten, daß ich vor der Benutzung des morphologischen Pendels die besten Trigramme von Hand auswählte.

Um wieder systematisch nach einer besten Einstellung für die Differenzanalyse der Trigramme zu suchen, variierte ich die Parameter systematisch und verzichtete dann auf eine manuelle Auswahl der besten Trigramme.

Das heißt: ich nahm *alle* Trigramme, die bei der Differenzanalyse gefunden wurden, als domänenspezifisch an. Dann bestimmte ich alle Wörter, die eines dieser Trigramme enthielten und verglich sie mit den Einworttermen, die der Experte identifiziert hatte. Daraus berechnete ich Precision, Recall und $F(0,5)$. Diesmal wurden aber von der Differenzanalyse (für Wörter) gefundene Terme nicht aus der Treffermenge eliminiert, so daß ein Recall bis zu 40% beobachtet wurde.

All dies führte ich einmal mit dem HQ und einmal mit der LR als statistischer Prüfgröße durch. Die drei Einstellungen, die den besten F-Wert für die beiden Texte und Prüfgrößen lieferten, sind in Tabelle 4.4 dargestellt.

	HQ		LR	
	(M; min_P)	F-Wert	(M; min_P)	F-Wert
Morphologie	(4; 18)	0,47	(1; 70)	0,41
	(4; 20)	0,47	(1; 68)	0,40
	(4; 22)	0,47	(9; 64)	0,40
Medizin	(2; 22)	0,59	(1; 70)	0,41
	(1; 22)	0,58	(1; 72)	0,41
	(2; 14)	0,66	(3; 70)	0,41

Tabelle 4.4: Paare (Mindestanzahl Wörter M; Schwellenwert Prüfgröße), für die $F(0,5)$ optimal ist

Es fällt auf, daß die LR diesmal bei beiden Texten schlechtere Ergeb-

nisse liefert als der HQ. Vielleicht kann man das damit erklären, daß domänenspezifische Trigramme von Natur aus „seltene“ Ereignisse sind, die gerne überbewertet werden dürfen — sie kommen nur in wenigen domänenspezifischen Morphemen vor, heben sich aber trotzdem (in ihrer relativen Frequenz) scharf von der Allgemeinsprache ab.

Insgesamt scheint das morphologische Pendel für den medizinischen Text besser zu funktionieren als für den linguistischen. Das verwundert wenig, da in der Medizin mehr lateinische und griechische Morphologie verwendet wird als in der Linguistik.

Außerdem muß für den linguistischen Text eine höhere Mindestanzahl M angenommen werden: das liegt — wie ich nach einigen Experimenten herausfand — daran, daß der linguistische Text eine Fülle von Beispielen z.B. aus afrikanischen Sprachen enthält. Setzt man beispielsweise die Mindestanzahl $M = 2$, so wird unter anderem das Trigramm „okm“ extrahiert, das in den Wörtern „chokma“ und „ikchokmo“ vorkommt. Diese sind natürlich keine linguistische Terminologie. Da aber nur wenige dieser Beispiele dasselbe Trigramm enthalten, kann man diese Trigramme durch das Heraufsetzen von M eliminieren, übrig bleiben Trigramme wie „rph“, das in vielen Termini („homomorph, morpheme, morphophonology, allomorph, ...“) vorkommt.

Wieder hat es sich also als unmöglich herausgestellt, eine global optimale (also für alle Texte gültige) Einstellung zu finden. Immerhin scheint der Schwellenwert für den HQ bei 22 für beide Texte eine gute Wahl zu sein. Ob man das allerdings auf weitere Texte verallgemeinern kann, ist fraglich.

Abbildung 4.6 zeigt den F-Wert $F(0,5)$ als Funktion der Mindestanzahl M und des Schwellenwertes für den HQ für den medizinischen Text. Man sieht eine gewisse Ähnlichkeit zu Abb. 4.3: bei der Differenzanalyse für Trigramme ergeben sich qualitativ in etwa dieselben Kurvenverläufe wie bei der Differenzanalyse für Wörter.

Um einen Einblick in die Arbeitsweise des morphologischen Pendels zu geben, möchte ich schließlich noch einige Trigramme präsentieren, die mit der optimalen Einstellung $(2; 22)$ im medizinischen Text gefunden wurden. Tabelle 4.5 zeigt die 10 Trigramme mit dem höchsten HQ-Wert zusammen mit jeweils zwei Beispielen. In Klammern hinter den Beispielwörtern steht dabei deren Frequenz im Text.

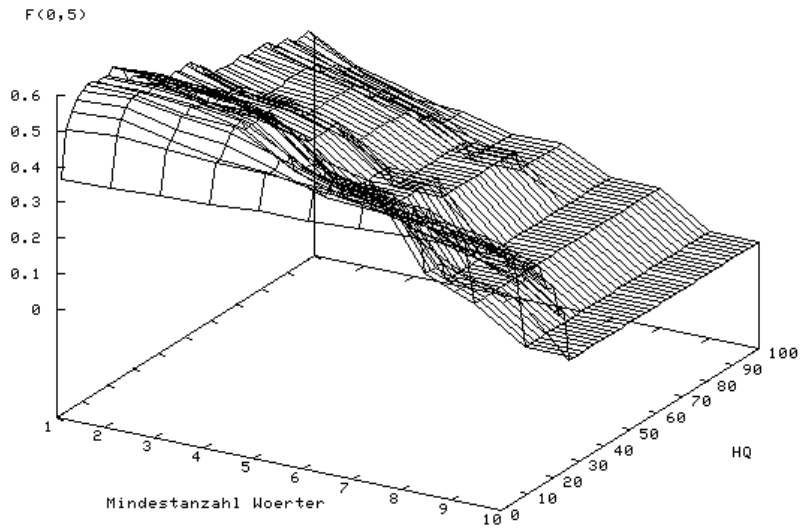

Abbildung 4.6: $F(0,5)$ als Funktion des HQ und der Mindestanzahl M (medizinischer Text)

Trigramm	Wörter, die das Trigramm enthalten (Frequenz)
bno	abnormality (27), abnormal (2)
stt	posttransplant (2), posttraumatic (2)
apy	chemotherapy (19), radiotherapy (1)
ysp	dysphagia (2), dyspnea (1)
hyr	triiodothyronine (1), thyroxine (1)
oim	autoimmune (1), autoimmunity (3)
cyt	thrombocytopenia (1), antithymocyte (1)
ocy	thrombocytopenia (1), antithymocyte (1)
pht	diphtheria (2), ophthalmologic (1)
oxa	dexrazoxane (1), cytoxan (2)

Tabelle 4.5: Gefundene Trigramme mit Beispielwörtern und deren Frequenz

Es fällt auf, daß ein großer Teil der so gefundenen Wörter eine Frequenz von 1 hat. Das ist der große Vorteil des morphologischen Pendels: es findet seltene Terme und ergänzt so sehr gut die Differenzanalyse: diese hat ja — gerade beim medizinischen Text, wie wir gesehen haben — Probleme, seltene Terme zu finden.

4.3.3 Syntaktisches Pendel

Für die Default-Einstellungen liefert das syntaktische Pendel beim medizinischen Text 43% Precision und 33% Recall, beim linguistischen ist $P = 37\%$ und $R = 48\%$ (der Recall bezieht sich jeweils nur auf die vom Experten identifizierten Phrasen, nicht aber auf Einwortterme).

Diese Ergebnisse sind zunächst enttäuschend. Darum möchte ich im Folgenden wieder untersuchen, wie sich das System bei Variation des Schwellenwertes für den C-Value verhält.

Variation des C-Value

Variiert man den Schwellenwert des C-Value systematisch in Schritten von 0,5 von 0 bis 10, so ergibt sich die in Abb. 4.7 gezeigte Kurve (für den medizinischen Text).

Das Optimum bezüglich des F-Wertes liegt hier bei einem Schwellenwert von 2,5 bzw. 3. Für den linguistuischen Text ist die Kurve leicht nach rechts verschoben, das Optimum liegt bei einem C-Value von 3,5 bzw. 4. Dieses inzwischen schon vertraute Phänomen erklärt sich wieder aus den oben genannten Überlegungen: im linguistischen Text wird kein Hintergrundwissen vorausgesetzt, die Phrasen werden — genauso wie die Einwortterme — öfter wiederholt als im medizinischen Text.

Einfluß des Termhood-Maßes

In Abschnitt 3.3.2 hatte ich als eine Verbesserung des Maßes C-Value die sogenannte *phrase_termhood* eingeführt, mit der jeder C-Value multipliziert wird: eine Phrase wird besser bewertet, wenn sie bereits bekannte Einwortterme enthält.

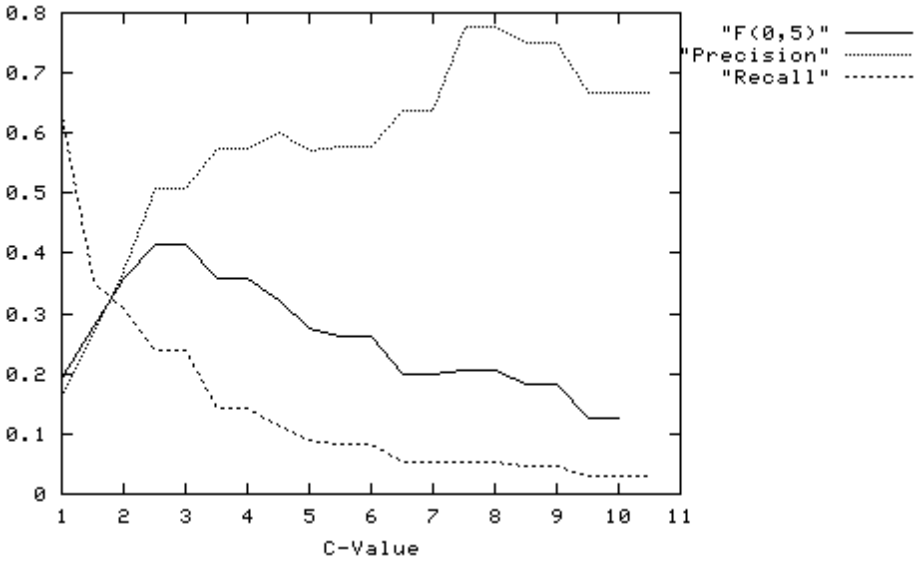

Abbildung 4.7: Precision, Recall und $F(0,5)$ bei Variation des C-Value

Abbildung 4.8 zeigt den Verlauf des F-Wertes einmal ohne und einmal mit Verwendung von *phrase_termhood*.

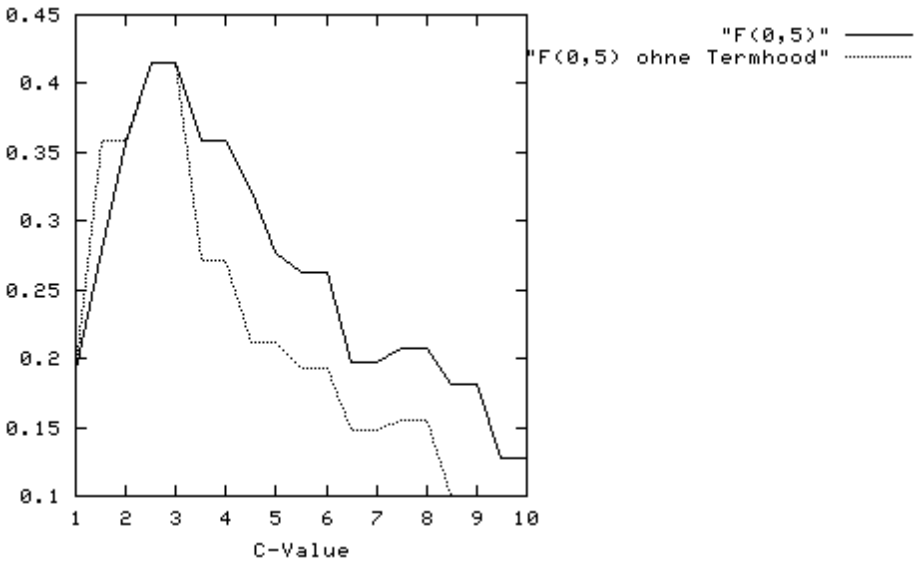

Abbildung 4.8: $F(0,5)$ als Funktion des C-Value mit und ohne *phrase_termhood*

Es fällt auf, daß für kleine Schwellenwerte *phrase_termhood* mehr scha-

det als nützt: die F-Werte sind für einen Schwellenwert von 1 und 1,5 ohne *phrase_termhood* besser als mit derselben.

Das ist natürlich erstaunlich, aber wie z.B. [Justeson 1995] feststellt: eine Phrase ist nur dann terminologisch, wenn sie *mehrmals* unvariiert auftritt. Durch die Verwendung der *phrase_termhood* erhalten manche Phrasen, die nur einmal auftreten, einen C-Value von 1,5 oder 2, weil sie einen bekannten Einwortterm enthalten. Somit überschreitet ihr C-Value womöglich den Schwellenwert, was aber nicht sinnvoll ist für dermaßen seltene Phrasen.

Das muß aber kein Problem sein, wenn man den Schwellenwert für den C-Value hoch genug ansetzt (man sieht ja, daß die F-Werte ab einem Schwellenwert von 2,5 mit *phrase_termhood* besser sind als ohne).

C-Value vs. Kollokationsmaß

Ich habe mich für das Maß C-Value entschieden, weil mir die Argumentation von [Justeson 1995] einleuchtete, die davon ausgeht, daß die pure Frequenz einer Phrase das beste Maß für deren Fachrelevanz ist.

Allerdings gibt es auch viele andere Ansätze, von denen die meisten die *Unithood* der Phrasen messen, unter Verwendung irgendeines Kollokationsmaßes (z.B. [Damerau 1993]). Das heißt, man möchte nicht nur wissen, wie oft zwei Wörter zusammen vorkommen, sondern auch, wie erstaunlich es ist, die Konstituenten der Phrase so oft nebeneinander zu sehen, wenn man von ihrer statistischen Unabhängigkeit ausgeht.

Um beide Ansätze vergleichen zu können, habe ich die Phrasenextraktion auch mit dem Poisson-Kollokationsmaß implementiert (vgl. Abschnitt 3.3.3 für die Formel). Das heißt eine Phrase wird extrahiert, wenn die Kollokationsstärke (die *Signifikanz*) zwischen den Konstituenten der Phrase einen gewissen Schwellenwert überschreitet.

Abbildung 4.9 zeigt den F-Wert in Abhängigkeit des Schwellenwertes für die Signifikanz, ebenfalls für den medizinischen Text.

Der maximale F-Wert liegt hier bei 0,35; mit dem C-Value erreicht man einen F-Wert von 0,42. Beim linguistischen Text bietet sich das gleiche Bild: mit Kollokationsmaß erreicht man dort $F = 0,28$, mit C-Value steigt F bis auf 0,54, d.h. hier ist der Unterschied sogar noch größer.

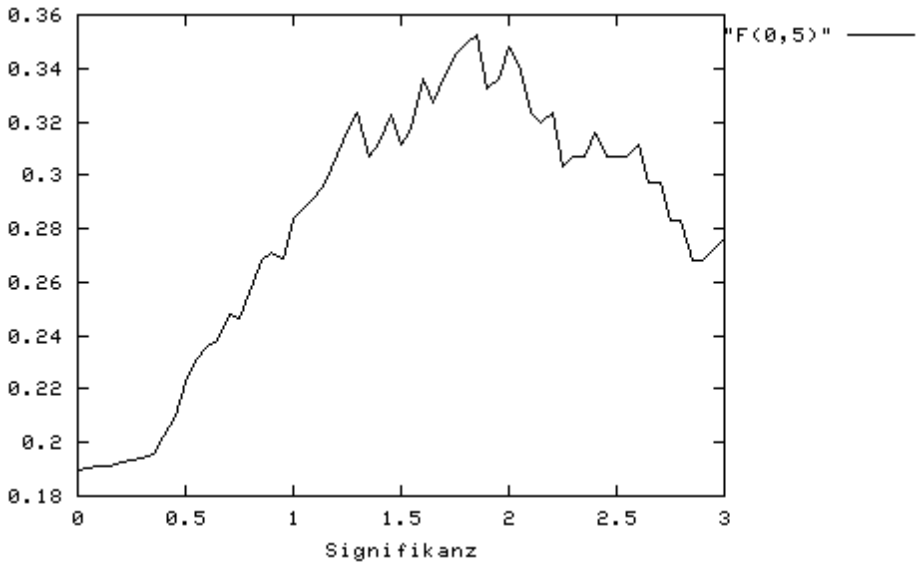

Abbildung 4.9: $F(0,5)$ als Funktion der Kollokationssignifikanz

Es fällt außerdem auf, daß die F-Wert-Kurve im Gegensatz zu Abb. 4.7 mehrere lokale Optima hat. Das macht es schwerer, nach einem optimalen Schwellenwert zu suchen.

Insgesamt scheint also die Wahl des C-Value — sofern man das nach der Betrachtung von zwei Texten sagen kann — keine schlechte gewesen zu sein: die Ergebnisse sind insgesamt besser als mit dem Kollokationsmaß.

Vergleich der POS-Muster

Nun möchte ich noch die Precision und den Recall der einzelnen POS-Muster untersuchen, um zu sehen, welches der Muster die besten Ergebnisse bringt bzw. welches man vielleicht lieber weglassen sollte.

Außerdem will ich die gefundenen Zahlen mit denen von [Arppe 1995] vergleichen.

Tabelle 4.6 zeigt die Ergebnisse für die Default-Schwellenwerte des C-Value für beide Texte.

Zunächst ist anzumerken, daß die Muster der Länge drei im medizinischen Text nur 14 Phrasen, im linguistischen Text sogar nur eine einzige Phrase fanden. Die Werte dieser Muster sind also wenig aussagekräftig.

	A N	N N	A A N	A N N	N N N
Morphologie	$P = 39\%$	$P = 39\%$	$P = 0\%$	$P = 0\%$	$P = 0\%$
	$R = 40\%$	$R = 8\%$	$R = 0\%$	$R = 0\%$	$R = 0\%$
Medizin	$P = 43\%$	$P = 46\%$	$P = 40\%$	$P = 43\%$	$P = 0\%$
	$R = 17\%$	$R = 12\%$	$R = 2\%$	$R = 2\%$	$R = 0\%$

Tabelle 4.6: Precision und Recall für die einzelnen POS-Muster

Weiterhin fällt auf, daß bei Arppes Mustern (siehe Tabelle A.1 im Anhang) auch etliche vorkommen, welche Präpositionen enthalten. Diese habe ich in den Default-Einstellungen nicht verwendet, vom Experten wurden aber in beiden Texten insgesamt nur zwei Phrasen mit Präposition identifiziert: die Begriffe „part of speech“ und „adverb of manner“ im linguistischen Text.

In anderen Fachdomänen gibt es vermutlich mehr Phrasen mit Präpositionen (z.B. Jura), für diese beiden Texte war aber die Entscheidung, auf POS-Muster mit Präpositionen zu verzichten, eindeutig richtig. Wieder stellt sich die Frage, ob das System selbst erkennen kann, wann solche POS-Muster lohnend sind oder ob der Anwender dies wissen muß.

Schließlich ist noch interessant, daß im Gegensatz zu Arppes Ergebnissen die Precision des Musters A N in beiden Texten nicht wesentlich unter der des Musters N N liegt. Auch wenn man den Schwellenwert des C-Values variiert, bleibt die Precision bei beiden Mustern ungefähr gleich groß. Der Recall des A N-Musters ist jedoch wesentlich höher als der von N N, was allerdings nach Arppes Ergebnissen auch zu erwarten war.

Insgesamt wurden aber wohl zu wenig Phrasen identifiziert (103 im medizinischen Text, 51 im linguistischen), um wirklich repräsentative Aussagen machen zu können.

4.3.4 Semantisches Pendel

Das semantische Pendel liefert für die Default-Einstellungen beim medizinischen Text 50 neue Wörter, beim linguistischen 14. Die Precision liegt im ersten Fall bei 62% (d.h. 31 der gefundenen 50 Wörter sind Fachterme), im

zweiten bei 43% (6 neue Terme). Der Recall ist in beiden Fällen gering; da das semantische Pendel als Ergänzung des Systems gedacht ist, stellt das aber kein Problem dar.

Noch einmal zur Erinnerung: die Defaulteinstellungen bestehen in diesem Falle darin, daß vor Anwendung des semantischen Pendels die Differenzanalyse mit den Einstellungen ($HQ = 1000$, Mindestfrequenz 2) und das morphologische Pendel mit den von Hand gewählten Trigrammen ausgeführt wird.

Alle dabei gefundenen Terme galten als „Input“ des semantischen Pendels, d.h. zu diesen wurden ähnliche gesucht. Dabei wurden jeweils so viele Iterationsschritte ausgeführt, bis Konvergenz eintrat (s.u. Abschnitt „Iteration“).

Wieder soll geklärt werden, ob sich das Ergebnis durch Variation von Parametern verbessern läßt. Dazu muß man die beiden Teile des Pendels getrennt untersuchen.

Aufzählungen

Auch bei den durch Aufzählungen gefundenen Termen hatte ich überlegt, ob es sich lohnt, einen Mindestwert der statistischen Prüfgröße als Filter zu benutzen. Variiert man diesen Schwellenwert so wie in den vorangegangenen Untersuchungen, so stellt man allerdings fest, daß er mehr schadet als nützt: der beste F-Wert ergibt sich, wenn man den Schwellenwert auf 0 setzt, ihn also ganz wegläßt.

Die Aufzählungen liefern 13 Wörter beim medizinischen Text (davon 9 Termini), wenn man auf einen Schwellenwert für die Prüfgröße verzichtet. Beim linguistischen Text werden 6 Wörter (davon 3 Termini) gefunden. Der Recall ist hier also noch geringer als beim Kollokationspendel.

Interessant fand ich noch, welche der in Abschnitt 3.3.3 vorgeschlagenen POS-Muster für Aufzählungen die meisten neuen Terme liefern. Für den linguistischen Text wurden *alle* neuen Termkandidaten mittels Aufzählungen der Form N C N (also so etwas wie „form and meaning“) gefunden, beim medizinischen waren es 10 der 13 neuen Kandidaten. Die restlichen 3 Wörter, die beim medizinischen Text gefunden werden, liefert das Muster

N C A N (Beispiel: „fever and other symptoms“). Die Muster, welche drei Nomina enthalten (N , N C N und N , N C A N), steuerten nichts bei.

Das liegt vermutlich daran, daß *alle* bis auf eines der beteiligten Nomina bereits Termstatus haben müssen, damit das fehlende diesen ebenfalls erhalten kann. Sind nun drei Nomina beteiligt, so müssen zwei dieser drei Termstatus haben, was offensichtlich selten auftritt. Diese strenge Handhabung ist aber notwendig, um eine gute Precision bei den Ergebnissen zu sichern.

Kollokationspendel

Bei der Suche nach ähnlichen Wörtern zu bereits gefundenen Einworttermen ist eine wichige Entscheidung, wie man den Schwellenwert für die Ähnlichkeit zweier Kollokationsvektoren festlegt, ab dem zwei Wörter als ähnlich gelten.

Ein weiterer Parameter ist der Schwellenwert für die statistische Prüfgröße der Differenzanalyse, den ich als einen Filter für die Ergebnisse benutzte. Dieser muß natürlich — damit man überhaupt noch etwas Neues findet — unter dem Schwellenwert der vorangegangenen Differenzanalyse liegen. Eine Mindestfrequenz gibt es beim semantischen Pendel aber nicht (es werden tatsächlich auch einige wenige Terme mit Frequenz 1 gefunden).

Abbildung 4.10 zeigt den F-Wert als Funktion der Mindestähnlichkeit (*Similarity*) und des Schwellenwertes der Prüfgröße HQ für den medizinischen Text.

Man sieht, daß offensichtlich der F-Wert mit steigendem Schwellenwert der *Similarity* abnimmt. Für den linguistischen Text bietet sich im Wesentlichen das gleiche Bild. Das ist sehr überraschend, denn es bedeutet, daß es offensichtlich optimal ist, *alle* Terme zu extrahieren, die mit einem der bereits gefundenen Terme eine Ähnlichkeit aufweisen, die größer ist als 0.

Der Schwellenwert des HQ hat offensichtlich einen größeren Einfluß: das Optimum liegt für den medizinischen Text bei $HQ = 100$, beim linguistischen bei $HQ = 40$. Das ist in beiden Fällen grob die Hälfte des optimalen Schwellenwertes bei der Differenzanalyse (siehe Tabelle 4.3).

Nun fragt sich, ob das semantische Pendel nicht einfach zu einer verkappten Differenzanalyse geworden ist, d.h. inwiefern sich die Ergebnisse

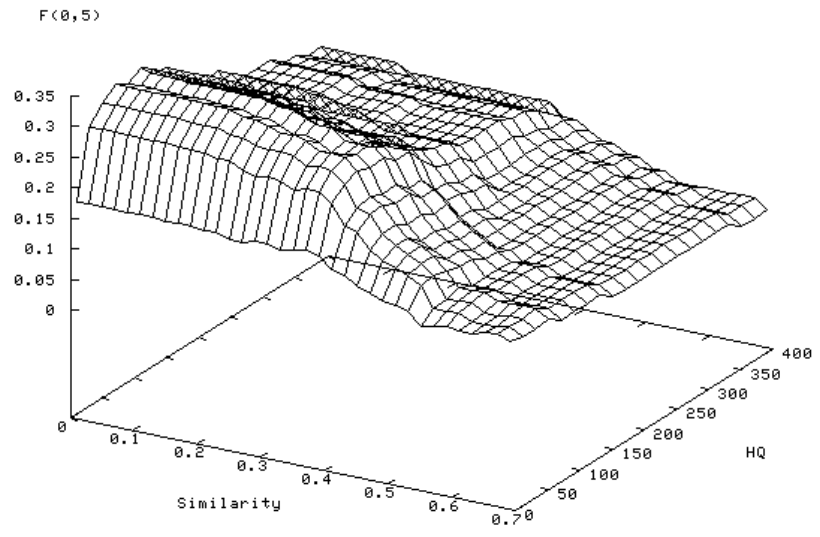

Abbildung 4.10: $F(0,5)$ als Funktion der Schwellenwerte von Similarity und HQ

des semantischen Pendels von einer Differenzanalyse unterscheiden, die als Schwellenwert $HQ = 100$ bzw. $HQ = 40$ hat.

Führt man für den medizinischen Text eine Differenzanalyse mit den Parametern ($HQ = 100$, Mindestfrequenz 1) durch, so ergeben sich — zusätzlich zu den Termen, die durch die Default-Einstellung ($HQ = 1000$, Mindestfrequenz 2) gefunden werden — 215 neue Termkandidaten. Das semantische Pendel findet nur 63 Wörter (die natürlich alle in der Menge der 215 Kandidaten der Differenzanalyse enthalten sind). Es gibt also einen deutlichen Unterschied zwischen semantischem Pendel und einer einfachen Differenzanalyse mit niedrigerem Schwellenwert.

Berechnet man jedoch die Precision für die 215 Terme, die mit der Differenzanalyse gefunden werden, so stellt man fest, daß diese mit 54% auch nicht wesentlich unter der des semantischen Pendels (68% mit optimalen Einstellungen) liegt. Der Recall der Differenzanalyse wäre hingegen mit 40%

wesentlich höher als der des semantischen Pendels (15%). Das heißt, es ist kaum ein Vorteil des semantischen Pendels gegenüber einer Differenzanalyse mit gesenktem Schwellenwert auszumachen.

Da auch der Rechenaufwand des semantischen Pendels sehr hoch ist gegenüber der Differenzanalyse, würde es vermutlich ausreichen, das semantische Pendel auf Aufzählungen zu reduzieren bzw. ganz wegzulassen. Möchte der Anwender den Recall erhöhen, so ist es wohl am sinnvollsten, wenn er einfach den Schwellenwert der Differenzanalyse heruntersetzt.

Iteration

Schließlich soll noch untersucht werden, wie sich das semantische Pendel bei mehrmaliger Iteration verhält.

Um dies herauszufinden, habe ich nach dem ersten Pendelschritt manuell die richtigen Fachtermini ausgewählt. Dann beobachtete ich, welche neuen Kandidaten im zweiten Pendelschritt gefunden wurden, wählte aus usw. Außerdem verwendete ich jeweils die oben gefundenen optimalen Einstellungen für das semantische Pendel.

Die Ergebnisse sind in Tabelle 4.7 dargestellt.

	Iterations-schritt	Gefundene Wörter	Davon Terme	Precision	Recall
Morphologie	1	23	11	48%	12%
	2	6	2	33%	2%
	3	1	0	0%	0%
Medizin	1	53	37	70%	13%
	2	10	6	60%	2%
	3	0	0	0%	0%

Tabelle 4.7: Entwicklung von Precision und Recall bei Iteration des semantischen Pendels

Bei beiden Texten kann man also offensichtlich nach dem zweiten — wenn nicht gar nach dem ersten — Pendelschritt aufhören.

Auch das ist gewissermaßen enttäuschend, da die Idee des „Pendelns“

ja davon ausgeht, daß man eine beträchtliche Zahl von Iterationsschritten ausführt, bis das Verfahren konvergiert.

4.4 Typische Probleme

Welche typischen Probleme verschlechtern Precision und Recall? Zunächst fällt bei der Betrachtung der Ergebnisse — beispielsweise der Differenzanalyse — auf, daß sich selbst bei niedrigsten Schwellenwerten kein erschöpfender Recall einstellt.

Das kann im Wesentlichen an vier Dingen liegen:

- Fehler bei der Grundformenreduktion: manche Wörter, wie z.B. „meaning“ (im Sinne von Bedeutung) werden vom Porter-Algorithmus falsch lemmatisiert (hier: „mean“) und können so niemals gefunden werden.
- Fehlende POS-Muster: manche Phrasen, die vom Fachmann identifiziert werden, haben sehr „exotische“ POS-Muster: z.B. hat der Begriff „graft versus host disease“ die Länge vier, was bei den zur Extraktion verwendeten Mustern gar nicht vorgesehen ist.
- Fehler bei der Wortsegmentierung: manche Wörter enthalten seltsame Zeichen. Da ich bei der Wortsegmentierung sogar Ziffern innerhalb von Wörtern verbiete, können Wörter wie „cd4“ (med. Text) nie gefunden werden. Ziffern zu verbieten ist aber trotzdem sinnvoll, da sonst die Precision mitunter sehr leidet.
- Falsche Einordnung als Eigennamen: manche Wörter, die vom POS-Tagger immer fälschlicherweise als Eigennamen getaggt wurden, werden als solche eingeordnet und somit nicht extrahiert.

Ein weiteres Problem für den Recall sind die bereits angesprochenen seltenen Terme. Beim medizinischen Text konnte man sehen, wie schwierig es ist, wenn ein großer Teil der Fachtermini nur einmal im Text vorkommt.

Die Precision kann hingegen durch „Verunreinigungen“ des Textes beeinträchtigt werden: im linguistischen Text kommen sehr viele Beispiele aus

verschiedenen Fremdsprachen vor — Wörter wie „wikiwiki“ oder „chokma“. Diese sind exotisch und fallen daher bei einer Differenzanalyse stark auf, obwohl sie keine Terminologie sind. Texte, in denen mehrere Sprachen gemischt vorkommen, stellen allgemein ein Problem für die Sprachverarbeitung dar.

Im Folgenden sollen noch die in Abschnitt 1.4.2 erwähnten Probleme untersucht werden:

Formale Variation

Die formale Variation — also das Auftreten eines Konzeptes in abgewandelter Form, wie „book review“ und „review of books“ — tritt in beiden von mir untersuchten Texten kein einziges Mal auf, d.h. weder der Experte noch LOGOS identifizierten irgendein fachsprachliches Konzept, welches in verschiedener Gestalt im Text auftrat. Es kann also kein großer Verlust sein, die formale Variation zu ignorieren.

Trennbarkeit von Phrasen

Alle vom Experten in beiden Texten identifizierten Phrasen sind — ohne Ausnahme — Nominalphrasen. Nominalphrasen sind nicht trennbar, das gilt für das Englische wie für das Deutsche gleichermaßen.

Das einzige Problem, das ich beobachtete, waren Konstruktionen wie „skin and breast cancer“. In dieser Phrase ist der Term „skin cancer“ enthalten, allerdings auseinandergerissen, so daß er mit dem POS-Muster N N nicht gefunden werden kann.

Dieses Problem trat in beiden Texten insgesamt viermal auf; es ist also ebenfalls hinreichend selten, um ignoriert werden zu können.

4.5 Vergleich mit anderen Systemen

Um meine Daten mit den Ergebnissen der Ansätze aus Kapitel 2 vergleichen zu können, suchte ich in den zugehörigen Veröffentlichungen nach Evaluierungen. Leider wurde ich nur bei dreien fündig. Die meisten IR-Ansätze enthalten zwar eine Evaluierung, jedoch werden nicht die gefundenen Index-

terme direkt bewertet, sondern ihr Einfluß auf das Auffinden von Dokumenten.

Die drei Veröffentlichungen, die eine Evaluierung enthielten, konzentrieren sich allesamt auf die Extraktion von Mehrwortbegriffen:

Bei [Damerau 1993] wurde eine Rangliste von Phrasen erstellt und dann mittels eines statistischen Tests das „Erstaunen“ darüber gemessen, so viele (oder wenige) fachspezifische Phrasen ganz oben in der Liste zu finden. Das Ergebnis des statistischen Tests wurde als Gütemaß verwendet. Dies läßt sich leider nicht mit meinen Precision-Recall-Experimenten vergleichen.

[Daille 1994] enthält eine Untersuchung der Precision: Phrasen werden mit zwei Kollokationsmaßen bewertet (Mutual Information (MI) und Likelihood Ratio) und die Precision für verschiedene Schwellenwerte abgetragen. Das entspricht genau meinem Ansatz, den Schwellenwert für den C-Value zu variieren. Die maximal erreichte Precision ist knapp 0,8 im Falle der LR, 0,5 im Falle der MI. Leider fehlt aber jegliche Aussage über den Recall bei diesen Precision-Werten.

Das ist das Problem beider bisher erwähnter Ansätze: es werden nur die generierten Listen betrachtet und daraus die Qualität der Ergebnisse, d.h. nur die Precision bewertet. Will man den Recall messen, so muß man wissen, wieviele und welche Fachtermini tatsächlich in dem jeweiligen Text enthalten sind. Dazu muß man den Text lesen, was natürlich sehr aufwendig ist.

Bei [Justeson 1995] wurde diese Arbeit für einen kurzen Text (noch kürzer als meine beiden Texte) durchgeführt. Ein Experte studierte den ganzen Text und annotierte alle dabei gefundene Terminologie. Er identifizierte 97 terminologische Phrasen. Das Extraktionssystem suchte nach Mehrworttermen mit einer Mindestfrequenz von 2 und fand 39 Phrasen, davon waren 36 terminologisch. Das entspricht also einem Recall von 37% und einer Precision von 92%. Leider wurde die Mindestfrequenz nicht variiert, man weiß also nicht, wie sich die Werte entwickeln.

Die gefundenen Werte sind jedenfalls besser als die von meinem System erzielten; allerdings wurden bei [Justeson 1995] noch Experimente mit zwei anderen — längeren — Texten durchgeführt. Dabei wurde die Precision

anhand der generierten Listen gemessen und es zeigte sich, daß die Precision mit der Länge des Textes bzw. auch mit der Anzahl der gefundenen Termini abnimmt: für 292 gefundene Phrasen liegt sie nur noch bei 67% (was sich eher mit meinen Erfahrungen deckt). Man weiß allerdings leider wieder nicht, wie es dabei um den Recall bestellt ist.

Justeson stellt außerdem fest, daß sowohl Precision als auch Recall sich deutlich verbessern, wenn man als Fachtermini nur diejenigen Wörter eines Textes ansieht, die wirklich zum speziellen Thema des Textes gehören. Für den oben untersuchten medizinischen Text hieße das z.B. eine Konzentration auf die Terme, die wirklich zum Thema „Knochenmarkstransplantation“ gehören, nicht allgemein zum Fachgebiet der Medizin.

Es ist logisch, daß in diesem Fall die Qualität steigt: die Terme, die zum engeren Thema des Textes gehören, werden öfter wiederholt und lassen sich daher leichter extrahieren. Diesen Effekt konnte man ja beim Vergleich des medizinischen und des linguistischen Textes zur Genüge beobachten.

Insgesamt ist der Vergleich mit anderen Systemen aber wenig erhellend; es fehlen ausreichende Vergleichsdaten, insbesondere z.B. auch zu anderen Ansätzen mit Differenzanalyse.

4.6 Fazit

Bei der Analyse der Komponenten des Systems habe ich festgestellt, daß die Differenzanalyse wohl der wichtigste Bestandteil des Systems ist: sie bietet eine relativ hohe Abdeckung bei gleichzeitig recht guter Precision. Welche Prüfgröße man verwendet, hängt dabei davon ab, ob man Hintergrundwissen extrahieren möchte: der HQ ist dafür geeignet, da er seltene Ereignisse „überbewertet“; die LR taugt eher zum Selektieren der wirklich themabezogenen (also hochfrequenten) Terme.

Als Ergänzung der Differenzanalyse zum Auffinden weiterer Einwortterme ist das morphologische Pendel aufgrund seiner hohen Precision eine gute Variante; das semantische Pendel dagegen erfüllt die Erwartungen nicht und sollte besser weggelassen werden.

Bei der Phrasenextraktion stellte sich das Maß C-Value — im Gegensatz zu anderen Maßen — als eine gute Wahl heraus; die Ergebnisse des syn-

taktischen Pendels insgesamt sind allerdings gegenüber der Differenzanalyse eher enttäuschend.

Zusammenfassend läßt sich festhalten, daß es keine Werte der Systemparameter gibt, die für alle Texte global optimale Ergebnisse versprechen.

Insgesamt gilt: möchte man nur die für das engere Thema eines Textes relevante Terminologie extrahieren, so muß man mit hohen Schwellenwerten arbeiten und wird dabei recht gute Ergebnisse erzielen. Dies entspricht in etwa einer *Beschlagwortung* des Textes mit inhaltlich relevanten Termen.

Will man hingegen auch das *Hintergrundwissen* in Form seltenerer Terme extrahieren, so stößt man auf Probleme, die umso größer sind, je mehr solches Hintergrundwissen in dem Text vorausgesetzt wird. Das Heruntersetzen der Parameter erlaubt zwar, einen großen Teil der Hintergrund-Terminologie zu extrahieren, jedoch leidet dabei die Precision in erheblicher Weise. Von der Beschaffenheit des Textes hängt auch ab, wie sehr man die Parameter herabsetzen muß; auch hier gibt es also kein allgemeingültiges Rezept.

Es wäre interessant herauszufinden, ob es möglich ist, daß das System selbst merkt, welche Textart vorliegt — dann könnte man die Parameter dynamisch anpassen. Wir haben einige Anhaltspunkte gesehen, die einen Hinweis auf die Textart geben können.

Eigentlich müßte man aber eine viel größere Zahl von Texten untersuchen, um auch weitere — außer den hier beobachteten — Effekten zu verstehen: z.B. ist unklar, wie genau sich die Parameter ändern müssen, wenn wesentlich längere Texte untersucht werden, oder solche, bei denen gar keine lateinische oder griechische Morphologie vorkommt (bei denen also das morphologische Pendel versagt).

Sehr viele und sehr lange Texte zu lesen und alle enthaltene Terminologie zu annotieren, ist allerdings zu aufwendig. Die gefundene Terminologie nur mit einem (elektronisch verfügbaren) Terminologie-Wörterbuch abzugleichen, halte ich auch für problematisch: viele der im Text enthaltenen Termini werden oft nicht im Lexikon zu finden sein, Aussagen über Recall und Precision sind somit schwierig.

Die Frage, ob sich auch ohne eine exzessive Untersuchung möglichst vie-

ler und langer Fachtexte eine Strategie für die dynamische Anpassung der Parameter finden läßt, soll Gegenstand des letzten Kapitels sein. Im nun folgenden Kapitel möchte ich noch kurz darauf eingehen, wie das System für die deutsche Sprache angepaßt werden kann.

Kapitel 5

Deutsch

In diesem Kapitel soll kurz beschrieben werden, welche Änderungen an meinem System nötig waren, damit es auch für die deutsche Sprache funktionierte bzw. welche zusätzlichen Verbesserungen im Deutschen möglich sind.

Dazu muß man sich zunächst klarmachen, daß im Wesentlichen zwei Unterschiede zwischen Englisch und Deutsch eine Rolle spielen werden:

1. Flexionsmorphologie: im Deutschen gibt es eine viel größere Anzahl von Flexionssuffixen und Flexionsklassen bei Nomina und Verben als im Englischen. Auch Allomorphie ist ausgeprägter. Das bedeutet, daß die Grundformenreduktion im Deutschen komplexer ist.

2. Komposition: Semantisch komplexe Konzepte werden im Deutschen meist durch Komposita ausgedrückt, im Englischen hingegen durch entsprechende Phrasen (Bsp.: „Strahlentherapie“ vs. „radiation therapy“). Das bedeutet, daß im Deutschen das syntaktische Pendel weniger ergiebig bzw. präzise sein wird. Stattdessen lassen sich über eine Kompositazerlegung weitere — morphologisch komplexe — Einwortterme finden, die im Englischen in dieser Form nicht existieren.

Abgesehen davon ist mein System auch in einigen weiteren Punkten sprachabhängig. Das heißt: Um das System für das Deutsche (oder eine beliebige andere Sprache) mit der gleichen Funktionalität auszustatten wie im Englischen, sind zunächst Anpassungen an folgenden Systemkomponenten notwendig:

- Grundformenreduktion: statt einer einfachen regelbasierten Lemmatisierung der Wörter sind im Deutschen ausgefeiltere Verfahren notwendig. Generell ist die Grundformenreduktion für jede Sprache neu zu realisieren.
- POS-Muster: die Muster ändern sich von Sprache zu Sprache, das Verfahren an sich bleibt unverändert.
- Stopwortliste: es muß eine Liste von Stopwörtern der jeweiligen Sprache verwendet werden.
- Referenzkorpus: es muß ein geeignetes Vergleichskorpus gefunden und aufbereitet werden.

Ein mögliches Verfahren für die **Grundformenreduktion** im Deutschen soll in Abschnitt 5.1 beschrieben werden. Die Aufbereitung des Referenzkorpus besteht dann im Wesentlichen in der Lemmatisierung aller Wörter und Bestimmung der Frequenzen für alle Grundformen. Als Referenzkorpus wählte ich den „Deutschen Wortschatz“ der Universität Leipzig, eine Sammlung vieler verschiedener deutscher Zeitungsartikel, Romane und Dichtungen.

Die Anpassung des syntaktischen Pendels hängt davon ab, welche Arten von Nominalphrasen in der jeweiligen Sprache bevorzugt als Fachterminologie auftreten. Die POS-Muster, die ich für das Deutsche auswählte, stammen von [Heid 1998]: $N\ Det\ N^{Gen}$, d.h. ein Nomen gefolgt von einem Artikel und einem weiteren Nomen im Genitiv, $N\ P\ N$ (P = Präposition) und $A\ N$.

Wie erwartet liefern diese Muster eine sehr niedrige Precision, d.h. es wäre überlegenswert, auf das syntaktische Pendel für das Deutsche zu verzichten. Stattdessen soll in Abschnitt 5.2 beschrieben werden, wie mittels einer **Kompositazerlegung** weitere morphologisch komplexe Einwortterme (also Komposita) gefunden werden können.

5.1 Grundformenreduktion

Problemstellung

Wie schon erwähnt, stellt die Grundformenreduktion im Deutschen ein durchaus schwieriges Problem dar. Das Deutsche weist sowohl bei Verben als auch bei Nomina und Adjektiven eine große Anzahl von Flexionsklassen auf, d.h. ein und dieselbe grammatische Kategorie kann durch verschiedene Flexionssuffixe (also Allomorphe) markiert werden, je nachdem welcher Lexikoneintrag gerade betrachtet wird.

Eine Flexionsklasse ist eine Menge von Wörtern, deren Vollformen mit Hilfe der jeweils gleichen Suffixallomorphe gebildet werden. So wird der Nominativ Plural vieler deutscher Nomina z.B. durch Anhängen von „-en“ gebildet (Bsp.: Sitzung, Sitzungen); bei anderen Nomina wird hingegen „-e“ suffigiert (Tisch, Tische), dabei kann auch ein Umlaut auftreten (Flug, Flüge); bei wieder anderen Nomina wird der Plural durch ein Nullmorphem realisiert (Fenster, Fenster).

Die Liste läßt sich noch um einiges verlängern. Vom linguistischen Standpunkt muß man — um ein Wort richtig lemmatisieren zu können — dessen Flexionsklasse kennen. Die Zuordnung von Wörtern zu Flexionsklassen ist im Deutschen aber willkürlich, man spricht auch von lexikalischer Allomorphie. Das heißt, beim Lernen der deutschen Sprache muß man zu jedem Wort dessen Flexionsklasse mitlernen. Das gilt eigentlich auch für das Englische, nur daß dort die Anzahl der Flexionsklassen wesentlich kleiner ist, man dort also eher von „Ausnahmen“ sprechen kann.

Es stellt sich die Frage, ob es nicht doch Eigenschaften eines Wortes gibt, welche Rückschlüsse auf dessen Flexionsklasse zulassen. Dies ist entscheidend wichtig für die Erstellung eines Regelsystems: im Englischen hatten wir gesehen, daß Nomina, die auf „-y“ enden, den Plural mit „-ies“ bilden. Im Deutschen gibt es ähnliche Regeln: Der Plural eines Wortes mit der Endung „-ung“ ist fast immer „-ungen“ (Sitzung, Sitzungen).

Das Problem liegt nun darin, daß im Deutschen

- aufgrund der großen Zahl der Flexionsklassen sehr viele dieser Regeln erstellt werden müssen und

- für viele Wörter keine so klaren Regeln existieren wie im Beispiel der Wörter mit Endung „-ung“

Letzteres soll an folgendem Beispiel illustriert werden:

Kanten → Kante
Folianten → Foliant
Atlanten → Atlas

Alle drei Vollformen enden auf „-anten“, gehören jedoch offensichtlich verschiedenen Flexionsklassen an; ein Regelwerk, das sie aufgrund ihrer Suffixe korrekt den unterschiedlichen Klassen zuweist, muß also mindestens 6 Buchstaben vom Ende der Wörter abschneiden.

Im Falle von „Kanten“ bleibt dann aber nichts mehr übrig. Kaum ein Mensch würde sich ein solches Regelwerk ausdenken, da es schnell zu unübersichtlich wird.

Angesichts dieser Probleme erscheint es sinnvoll, eine Regelmenge nicht von Hand zu erstellen — man wird wohl kaum alle Besonderheiten bedenken können. Stattdessen bietet es sich an, ein System zu konstruieren, welches aus einer möglichst großen, korrekt lemmatisierten Trainingsmenge lernt und in der Lage ist, die dieser Trainigsmenge inhärenten Regelmäßigkeiten sinnvoll für unbekannte Wörter zu verallgemeinern.

Wählt man als Trainingsmenge eine Liste der häufigsten Wörter eines gemeinsprachlichen Korpus, so ist auch garantiert, daß häufige Phänomene korrekt behandelt werden. Bei seltenen Ausnahmen werden sich einige Fehler nicht vermeiden lassen.

Affix Compression Tries

Ein mögliches Lernverfahren für die automatische Grundformenreduktion wird in [Biemann 2004] (Abschnitt 4.1) beschrieben: sogenannte Affix Compression Tries (ACT). Diese haben eine gewisse Ähnlichkeit mit Entscheidungsbäumen — ein Wort soll aufgrund seiner Endung in die richtige Flexionsklasse eingeordnet werden. Die Flexionsklasse wird dabei durch eine Reduktionsregel repräsentiert.

Ein ACT enthält alle Wörter einer Trainingsmenge mit zugehörigen Reduktionsregeln. Tabelle 5.1 zeigt eine solche Trainigsmenge.

Wort	Regel
Haus	0
Hauses	2
Häuser	5aus
Bau	0
Baus	1
aus	0

Tabelle 5.1: Trainingsmenge zur Grundformenreduktion

Eine Regel besteht aus einer Zahl und optional einem Suffix. Die Zahl besagt, wieviele Buchstaben vom Ende des Wortes abgeschnitten werden müssen, das optionale Suffix muß danach evtl. wieder angehängt werden, um die Grundform zu erhalten.

Abbildung 5.1 zeigt den zugehörigen ACT: alle Wörter der Trainigsmenge sind zusammen mit ihren Regeln dort eingespeichert. Dabei wird das Wort von hinten gelesen, gemeinsame Buchstabenfolgen der Wörter werden in einem Knoten zusammengefaßt. Ist man in einem Blatt des Baumes angekommen, so hat man ein komplettes Wort (rückwärts) eingelesen.

Im unteren Teil eines Knotens K ist angegeben, welche Regeln bei den unterhalb von K abgespeicherten Wörtern vorkommen; in Klammern steht zu jeder Regel, wie oft sie zur Anwendung kam.

Hierzu ein Beispiel: Der Knoten mit dem Eintrag „s“ hat unter sich alle Wörter der Trainingsmenge, die auf „-s“ enden. Bei diesen kam einmal die Regel „2“ vor (d.h. schneide zwei Buchstaben ab, nämlich bei „Hauses“), einmal die Regel „1“ (bei „Baus“) und zweimal die Regel „0“ (bei „Haus“ und „aus“).

Möchte man nun ein Wort lemmatisieren, so sucht man es im ACT. Ist das Wort enthalten, so endet die Suche in einem Blatt des Baums; dort steht die richtige anzuwendende Regel.

Ist das Wort nicht enthalten, so endet die Suche, sobald kein Match mehr mit dem Suffix des Suchwortes auftritt, also meist bei einem inneren Knoten

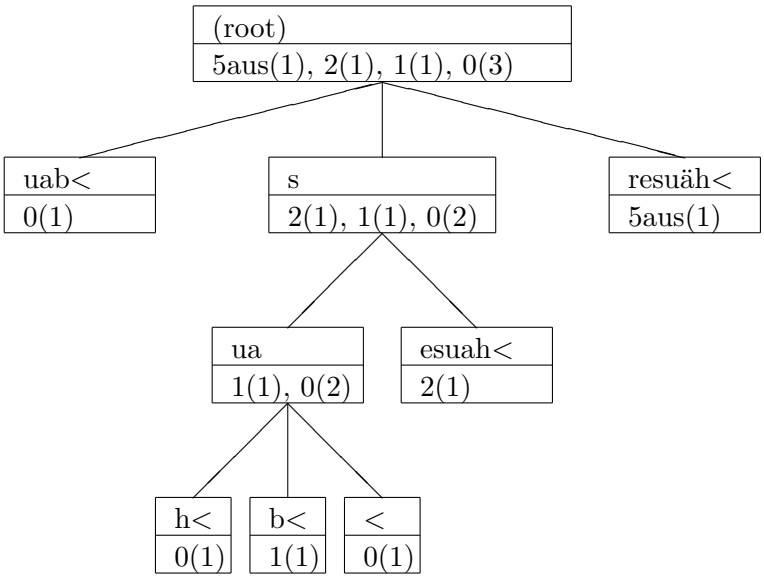

Abbildung 5.1: Beispiel für einen Affix Compression Trie

des Baums. Dann wird diejenige Regel ausgewählt, die in diesem Knoten am häufigsten vertreten ist.

Sucht man in unserem Beispiel-ACT nach dem Wort „Gans“, so endet die Suche im Knoten „s“. Die häufigste Regel dort ist „0“, das Wort „Gans“ wird also so belassen wie es ist. Sucht man hingegen nach „Hochhauses“, so landet man im Knoten „esuah<“. Die dort gespeicherte Regel besagt, daß die letzten zwei Buchstaben des Eingabewortes abgeschnitten werden sollen, man erhält also „Hochhaus“.

Das so entstandene Regelsystem hat zwei wichtige Eigenschaften:

1. Die Wörter der Trainingsmenge werden zu 100% korrekt lemmatisiert.

2. Trotzdem ist der ACT verallgemeinerungsfähig: ist ein Wort nicht enthalten, so wird eine Vorhersage getroffen, die auf den häufigsten Regeln beruht, die bei Wörtern mit gleicher Endung zur Anwendung kamen.

Das Verfahren funktioniert am besten, wenn man für jede der drei flektierenden Wortarten des Deutschen (Nomina, Verben und Adjektive) einen

eigenen ACT aufbaut und dann jedes Wort in dem entsprechenden ACT sucht.

Das ist sofort einleuchtend, wenn man sich z.B. überlegt, daß der Plural vieler Substantive durch Anhängen von „-en“ gebildet wird. Dort sollte „-en“ also abgeschnitten werden; bei Verben hingegen, deren Infinitiv als Grundform angesehen werden soll, ist das Abschneiden von „-en“ meist falsch (z.B. gehen → geh).

Ich benutzte eine an der Universität Leipzig verfügbare Implementierung der ACTs für meine Grundformenreduktion. Wie erwartet ist die Qualität nicht ganz so hoch wie im Englischen, jedoch nur unwesentlich.

Mit der Integration dieser Komponente war mein System also in der Lage, für die deutsche Sprache dasselbe zu leisten wie für Englisch. Im folgenden Abschnitt soll beschrieben werden, welche Verbesserungen darüber hinaus möglich sind, wenn man eine Kompositazerlegung miteinbezieht.

5.2 Kompositazerlegung

Wie bereits erwähnt, verspreche ich mir eine Verbesserung meines Systems durch eine Suche nach Komposita mit domänenspezifischen Morphemen. Dies scheint mir aus zwei Gründen vielversprechend:

1. Ein großer Teil der Terminologie deutscher Fachsprachen besteht aus Komposita.

2. Genauso wie in manchen Domänen bestimmte Derivationssuffixe besonders produktiv sind (siehe Abschnitt 1.4.1), so ist auch bei der Komposition zu beobachten, daß manche Basismorpheme in vielen verschiedenen fachspezifischen Komposita einer Domäne auftreten.

Unter einem domänenspezifischen Basismorphem möchte ich also im Folgenden ein Morphem verstehen, welches in vielen *verschiedenen* Komposita eines Fachtextes auftritt. Dem Ansatz von [Heid 1998] folgend, ergibt sich damit ein Algorithmus zur Extraktion solcher Morpheme, wie er in Tabelle 5.2 gezeigt wird.

```
1 Zerlege alle (verschiedenen) Wörter des Textes mit
   Hilfe eines Kompositazerlegers
2 Für jedes gefundene Kompositum X
      - bestimme dessen Teile X1,...,Xn
      - Zähle die Frequenz freq(Xi) jedes Teilmorphems
        Xi eins hoch
3 Gib alle Basismorpheme aus, für die freq(Xi)
   > Schwellenwert.
```

Tabelle 5.2: Algorithmus zum Auffinden domänenspezifischer Basismorpheme

Die so gefundenen Basismorpheme werden wieder dem Anwender zur Durchsicht vorgelegt; dabei wird als zusätzliche Information die Anzahl (verschiedener) Komposita präsentiert, in denen das Morphem enthalten ist, sowie drei dieser Komposita als Beispiele.

Der Benutzer entscheidet sich sodann für die jeweils besten Basismorpheme, und es werden — ganz analog zum Vorgehen bei Buchstabentrigrammen — alle Wörter extrahiert, die eines der gewählten Basismorpheme enthalten.

Die Suche nach Komposita ist also ein Teil des morphologischen Pendels, der vollkommen analog zur Suche mittels domänenspezifischen Buchstabentrigrammen verläuft, nur daß statt der Trigramme ganze Morpheme verwendet werden.

Es bleibt noch zu erklären, wie der Kompositazerleger funktioniert, den ich benutzte: es handelt sich auch hierbei um ein Lernverfahren mit ACTs. Als Trainingsmenge wird eine Liste deutscher Wörter verwendet; zu jedem Wort gibt es diesmal zwei Regeln, die jeweils besagen, wieviele Buchstaben des Eingabewortes von vorne bzw. hinten abzuschneiden sind, um die Teile des Kompositums zu erhalten.

Ein Beispiel: zum Wort „Hochgebirge“ gibt es die Regel „4“ (schneide von vorne „Hoch“ ab) und die Regel „7“ (schneide von hinten „gebirge“ ab).

Bei „Hochzeit“ gibt es hingegen zweimal die Regel „0“, d.h. das Wort soll gar nicht zerlegt werden.

Daraus werden zwei ACTs aufgebaut, in denen ein zu zerlegendes Wort gesucht wird. Nur wenn die Schnittstelle, die beide ACTs voraussagen, übereinstimmt, wird das Wort an dieser Stelle zerlegt.

5.3 Ergebnisse

Die Ergebnisse der deutschen Terminologie-Extraktion sollen an dieser Stelle nur ganz kurz und auch nur qualitativ betrachtet werden. Für die Untersuchung wählte ich einen Text, der dem medizinischen Text aus Kapitel 4 thematisch ähnlich ist: er beschreibt Diagnose- und Behandlungsmethoden für zwei verschiedene Formen der Leukämie. Die Länge des Textes beträgt 3316 Wörter.

Die Ergebnisse der Differenzanalyse für Wörter und Buchstabentrigramme und des semantischen Pendels unterscheiden sich — logischerweise — qualitativ nicht von denen des englischen Textes. Untersuchen möchte ich also vor allem die Verhaltensweise des syntaktischen Pendels und der neu eingeführten Kompositaextraktion.

Syntaktisches Pendel

Wie in der Einleitung dieses Kapitels angekündigt, wählte ich drei POS-Muster für das syntaktische Pendel im Deutschen:

$N\ Det\ N^{Gen}$
$N\ P\ N$
$A\ N$

Die Ergebnisse des letzten Musters sind mit denen des englischen Textes vergleichbar. Zwar gibt es im Deutschen auch eine relativ hohe Anzahl an Komposita, die aus einem Adjektiv und einem Nomen bestehen (z.B. „Hochdruck“ statt „hoher Druck“). Trotzdem werden viele Konzepte, die aus Adjektiv und Nomen bestehen, durch eine entsprechende Phrase realisiert (z.B. „kinetische Energie“ und nicht „Kinetikenergie“ oder Ähnliches).

Das erklärt, warum auch im Deutschen das Muster $A\ N$ relativ ergiebig ist.

Für die Kombination zweier Nomina wird im Deutschen jedoch fast ausschließlich die Komposition gewählt. Phrasen, die dem ersten oder zweiten der obigen POS-Muster entsprechen, sind selten und meist nicht terminologischer Natur.

Um dies zu illustrieren, zeigt Tabelle 5.3 alle Phrasen des untersuchten Textes, die mittels der beiden Muster extrahiert wurden (bei einem Mindest-C-Value von 2).

Muster	Phrasen, die dem Muster entsprechen
$N\ Det\ N^{Gen}$	Erreichen der Remission
	Therapie des Rezidivs
	Erfolgsaussicht einer Rezidivtherapie
	Durchführung einer Schädelbestrahlung
$N\ P\ N$	Nachweis von Lymphoblasten
	Ara-C in Kombination
	Blasten im Blut

Tabelle 5.3: Deutsche Phrasen aus einem medizinischen Text

Zum Vergleich: Das Muster $A\ N$ findet 25 Phrasen. Man sieht also, daß die Ergiebigkeit der beiden Muster gering ist. Darüberhinaus ist keine der gefundenen 7 Phrasen wirklich terminologisch.

[Heid 1998] schlägt daher vor, nur solche Phrasen aus zwei Nomina zu akzeptieren, bei denen beide Nomina bereits (nach der Differenzanalyse) Termstatus haben. Dies trifft für keine der obigen Phrasen zu.

Vielleicht war also der Text zu klein gewählt, um repräsentative Aussagen zuzulassen. Da mein System aber gerade für kleine Texte gute Ergebnisse liefern soll und die beiden Muster gegenüber dem Muster $A\ N$ dermaßen unergiebig sind, werde ich das syntaktische Pendel im Deutschen auf das Muster $A\ N$ beschränken.

Kompositaextraktion

Da das syntaktische Pendel sehr schlechte Ergebnisse liefert, soll hier die Behauptung geprüft werden, daß die Kompositaextraktion einen adäquaten Ersatz bietet.

Tabelle 5.4 zeigt die im deutschen Text nach der Kompositazerlegung gefundenen Basismorpheme. Die Zahl in Klammern besagt, in wievielen *verschiedenen* Komposita das Basismorphem auftritt; angezeigt werden nur Basismorpheme, die in mindestens vier Komposita vorkamen.

therapie (29), tisch (12), rezidiv (11), leukämie (10), risiko (9), knochen (7), mark (7), system (6), niere (6), organ (6), typ (6), tier (6), zyto (6), gisch (5), behandlung (5), infiltrat (5), remission (5), untersuchung (4), zahl (4), blutung (4), blut (4), diagnostik (4), chemo (4), dauer (4), hode (4), punktion (4), erkrankung (4)

Tabelle 5.4: Gefundene Basismorpheme

Man sieht, daß nicht alle der gefundenen Basismorpheme wirklich Basismorpheme im linguistischen Sinne sind (siehe Abschnitt 1.4.1): „tisch“ und „gisch“ bezeichnen keine Sachverhalte der außersprachlichen Welt, sie ähneln eher Derivativen. Das Morphem „tier“ fällt auch in diese Kategorie, da es aus Wörtern wie „akzeptieren“ gewonnen wurde.

Die Kompositazerlegung ist also nicht perfekt, was negative Auswirkungen auf die extrahierten Basismorpheme hat. Daher ist hier wieder die Intervention des Anwenders sinnvoll.

Tabelle 5.5 zeigt die Komposita, welche mit einer (von mir getroffenen) Auswahl der obigen Basismorpheme gefunden wurden. In Klammern steht diesmal die Frequenz des jeweiligen Kompositums im Text. Man sieht, daß auch hier wieder viele Einwortterme mit Frequenz 1 gefunden werden.

Ich denke, diese Ergebnisse bestätigen relativ eindrucksvoll die eingangs aufgestellte These: Das syntaktische Pendel kann im Deutschen durch die Kompositaextraktion fast vollständig und ohne Qualitätseinbußen ersetzt werden. Die Tatsache, daß viele der gefundenen Terme nur einmal im Text auftreten (und trotzdem terminologisch sind), stellt sogar eine Verbesserung

des Recalls gegenüber dem syntaktischen Pendel dar: dieses kommt nicht ohne Mindestfrequenz (bzw. C-Value) aus.

Insgesamt sind die Ergebnisse der deutschen Terminologie-Extraktion durchaus nicht schlechter als die der englischen. In beiden Sprachen gibt es aber noch eine Vielzahl von möglichen Verbesserungen und offenen Fragen, denen im letzten Kapitel nachgegangen werden soll.

Basismorphem	Komposita, die das Basismorphem enthalten
Rezidiv	Rezidivtherapie (3), Rezidivrisiko (3), Hodenrezidiv (2), ZNS-Rezidiv (2), Spätrezidiv (1), Knochenmark-Rezidiv (1), Rezidivwahrscheinlichkeit (1), Rezidivkaskade (1), Rezidivpatient (1), Rezidivverdacht (1), Rezidivbehandlung (1)
Leukämie	Leukämiezelle (4), Promyelozyten-Leukämie (2), Leukämiezellmaß (1), Erythroblasten-Leukämie (1), B–Zell-Leukämie (1), Leukämiebefall (1), Leukämiediagnostik (1), T-Zell-Leukämie (1), Monoblasten-Leukämie (1), Monozyten-Leukämie (1)
Knochen	Knochenmark (5), Knochenmarktransplantation (5), Knochenmarkpunktion (3), Knochenmark-Punktion (1), Knochenmark-Rezidiv (1), Knochenmarkaplasie (1), Knochenmarkinsuffizienz (1)
System	systemisch (3), Organsystem (1), Systemerkrankung (1), Nervensystem (1), Zentralnervensystem (1), Kathetersystem (1)
Organ	Organfunktion (2), Organinfiltrat (1), Organomegalie (1), Organsystem (1), Organgröße (1), Organtoxizität (1)
Typ	Phänotyp (4), Immunphänotyp (4), FAB-Typ (2), Burkitt-Typ (1), AML-Subtyp (1), FAB-Subtyp (1)
Zyto	Zytochemie (4), Zytogenetik (2), Zytostatikabehandlung (1), Zytostatikaresistenz (1), Impulszytophotometrie (1), zytoplasmatisch (1)
Infiltrat	Niereninfiltrat (2), Hautinfiltrat (1), Organinfiltrat (1), Skelettinfiltrat (1), Darminfiltrat (1)
Remission	Remissionsbeurteilung (2), Remissionskriteri (2), Remissionsüberwachung (1), Postremissionschemotherapie (1), Remissionsstatus (1)

Tabelle 5.5: Ergebnisse der Kompositaextraktion

Kapitel 6

Ausblick

In diesem Kapitel möchte ich Verbesserungen und Erweiterungen meines Extraktions-Systems beschreiben, die im Rahmen dieser Diplomarbeit nicht mehr realisiert werden konnten und somit Gegenstand zukünftiger Arbeit sein werden.

Dabei soll zunächst darauf eingegangen werden, welche Mängel am aktuellen System noch behoben werden müßten. Im Wesentlichen handelt es sich dabei um die in Kapitel 4 angesprochenen Probleme: es stellt sich die Frage, wie bzw. ob man die optimalen Parameter für die Extraktion automatisch anhand gewisser Eigenschaften eines Textes feststellen kann.

In den darauffolgenden Abschnitten sollen dann noch einige mögliche Erweiterungen des Systems vorgestellt werden, die über eine reine Extraktion von Fachtermini hinausgehen.

6.1 Dynamische Anpassung von Parametern

Wie in Kapitel 4 gesehen, hängt die optimale Einstellung vieler Parameter der Extraktion stark davon ab, welche Sorte von Text bearbeitet wird. Anhand des medizinischen und des linguistischen Textes wurde bereits deutlich, daß Text nicht gleich Text ist, sondern daß vielmehr jeder Text gemäß seiner individuellen Eigenschaften anders behandelt werden muß.

6.1.1 Textsorten

Zunächst muß man sich überlegen, welche Eigenschaften eines Textes einen Einfluß auf die Extraktionsparameter haben. Das heißt: welche Unterschiede zwischen zwei Texten machen eine unterschiedliche Behandlung der beiden nötig?

Die nun folgende Aufzählung ist nur eine Sammlung von Ideen zu diesem Thema und erhebt keinen Anspruch auf Vollständigkeit. In Kapitel 4 wurden bereits — anhand des medizinischen und des linguistischen Textes — zwei Dinge deutlich:

- Einführende vs. „Expertentexte“: Hierbei handelt es sich um die Frage, wieviel Hintergrundwissen in einem Text vorausgesetzt wird. Jeder Text vermittelt Wissen, oft auch indem neue Fachtermini eingeführt und erklärt werden. Ein einführender Text (wie der linguistische aus Kapitel 4) verwendet dazu keine weiteren — als bekannt vorausgesetzten — Fachtermini derselben Domäne. Ein „Expertentext“ hingegen (wie der medizinische) baut auf einem Vorwissen in Form domänenspezifischer Termini auf und führt weitere, speziellere ein.

  In „Expertentexten“ existieren also zwei Arten von Termen: solche, die neu eingeführtes Wissen repräsentieren und so etwas wie „Hintergrundterme“ (vgl. hierzu auch [Justeson 1995]). Hintergrundterme treten vermutlich eher selten auf (wie am medizinischen Text sichtbar wurde), während das neu eingeführte Wissen zum besseren Verständnis wiederholt wird.

  Will man also bei einem „Expertentext“ auch Hintergrundwissen extrahieren, so muß man die Schwellenwerte, z.B. die Mindestfrequenz von Termen, senken.

- Facettenreichtum: Wieviele verschiedene Facetten eines Themas bzw. wieviele verschiedene (Teil-)Themen werden beleuchtet? Wird in einem Text ein Überblick über verschiedene Teilbereiche eines Themas gegeben (wie z.B. im medizinischen Text aus Kapitel 4), so treten die Terme, die nur für einen dieser Teilbereiche wichtig sind, vermutlich seltener auf. Darüberhinaus clustern sie stark, d.h. sie treten nur in ei-

nem Absatz oder Kapitel des Textes gehäuft auf, außerhalb jedoch so gut wie gar nicht. In Texten, welche nur ein Thema behandeln, sind die Terme hingegen gleichmäßig über den Text verteilt und treten meist insgesamt mit einer höheren Frequenz auf.

Werden Teilbereiche behandelt, so existieren wieder zwei Sorten von Fachtermen: solche, die dem übergeordneten Thema des Textes angehören und solche, die nur für einen der Teilbereiche wichtig sind. Die „Teilbereichsterme“ clustern stark und sind insgesamt selten, die „übergeordneten“ Terme verteilen sich gleichmäßig und treten mit höherer Frequenz auf.

Wieder gilt also: will man „Teilbereichsterme“ extrahieren, so muß man die Schwellenwerte der Extraktionsparameter senken.

In Kapitel 4 konnte man beobachten, daß beide Phänomene zuweilen Hand in Hand gehen: einführende Texte sind oft monothematisch, Expertentexte behandeln manchmal mehrere Themen. Entsprechend ist in letzteren (also z.B. im medizinischen Text) die Menge der „Teilbereichsterme“ identisch mit dem Hintergrundwissen bzw. die übergeordneten Terme entsprechen dem eigentlichen Thema des Textes (hier: Knochenmarkstransplantation) und sind somit neu eingeführtes Wissen.

Nun gibt es aber noch weitere Eigenschaften eines Textes, die einen Einfluß haben auf die Extraktionsparameter:

- Domänenspezifische Morphologie: wie stark sich die Buchstabentrigramme eines Fachtextes von denen der Allgemeinsprache unterscheiden, hängt von der Fachdomäne ab: in Domänen wie Medizin, Chemie oder Biologie, in denen viel lateinische oder griechische Morphologie verwendet wird, hat eine Differenzanalyse für Trigramme einen wesentlich größeren Erfolg als in Domänen wie Jura, in denen nur mit Morphologie der jeweiligen Sprache gearbeitet wird.

  In juristischen Texten gibt es dafür aber eine wesentlich höhere Anzahl von Phrasen (im Englischen) bzw. Komposita (im Deutschen): z.B. Terme wie „unverschuldete Arbeitsunfähigkeit“ bzw. „incapacity

to work at no fault of the employee“.[1] Termini in den Domänen Medizin, Chemie oder Biologie sind hingegen sehr kompakte Einwortterme.

Wie stark sich die Trigramme abheben, sollte erkannt werden, damit entsprechend der Domäne ein Schwerpunkt auf Trigrammanalyse bzw. auf das syntaktische Pendel gelegt werden kann.

- Textlänge: ab einer bestimmten Länge eines Textes müssen sicherlich höhere Mindestfrequenzen für Terme verwendet werden.

- Eigennamen: will man Eigennamen extrahieren oder nicht? Auch dies ist abhängig von der Fachdomäne des Textes: in wissenschaftlichen Arbeiten treten Eigennamen fast nur in Zitaten bzw. im Literaturverzeichnis auf und sind somit uninteressant (eine Ausnahme bilden hierbei physikalische Einheiten wie z.B. „Hertz“, die sehr wohl extrahiert werden sollen).

  In Firmenberichten oder sonstigen Texten aus dem betriebswirtschaftlichen Umfeld treten Eigennamen hingegen oft als Bezeichner von Firmen oder Produkten auf. Diese sind potentiell interessant, sollten also extrahiert werden.

  Man müßte daher in der Lage sein, einen wissenschaftlichen Artikel automatisch von einem Firmenbericht zu unterscheiden und die Extraktion von Eigennamen entsprechend anzupassen.

6.1.2 Lösungsansätze

Die oben genannten Eigenheiten eines Textes automatisch erkennen und somit die Parameter anpassen zu können, ist sicherlich keine leichte Aufgabe. Hier sollen einige Vorschläge untersucht werden.

Ansätze in der Literatur

In der in Kapitel 2 eingeführten Literatur beschäftigen sich nur zwei Ansätze mit der dynamischen Anpassung von Parametern:

[1] siehe Terminologiesammlung aus Diplomarbeiten der FH Köln. URL: http://www.iim.fh-koeln.de/webterm/webtermsamm_d.htm, Abschnitt „Arbeitsvertrag“

1. In [Justeson 1995] wird vorgeschlagen, die Mindestfrequenz von Phrasen proportional zur Textlänge ansteigen zu lassen (Abschnitt 4.2). Dies löst zunächst das oben angesprochene Problem „Textlänge“.

   Die alleinige Verwendung dieses Ansatzes greift jedoch viel zu kurz: in Kapitel 4 mußte für den *kürzeren* der beiden Texte (den linguistischen) eine *höhere* Mindestfrequenz für Terme angesetzt werden. Das hatte mit den ersten beiden der oben aufgeführten Punkte zu tun, die Justeson ignoriert: der linguistische Text ist ein einführender, monothematischer Text, die verwendeten Fachtermini wiederholen sich also oft.

2. Eine dynamische Anpassung des Schwellenwertes der statistischen Prüfgröße einer Differenzanalyse wird von [Cohen 1995] vorgeschlagen: Er benützt als Schwellenwert den Mittelwert plus ein Vielfaches der Standardabweichung der statistischen Prüfgröße aller n-gramme.

Ich habe versucht, Cohens Idee auf die Differenzanalyse von Termen zu übertragen: Berechnet man für den medizinischen und den linguistischen Text jeweils den Mittelwert der likelihood ratio für alle Terme und die zugehörige Standardabweichung, so erhält man die Ergebnisse in Tabelle 6.1.

	Mittelwert	Standardabweichung
medizinischer Text	19,2	60,7
linguistischer Text	15,0	87,6

Tabelle 6.1: Mittelwerte und Standardabweichungen der LR für alle Wörter beider Texte

Es fällt auf, daß die Standardabweichung in beiden Fällen wesentlich größer ist als der Mittelwert. Die Werte der LR sind also sehr stark gestreut. Cohen berechnet den Schwellenwert s als Mittelwert m plus zweimal Standardabweichung σ:

$$s = m + 2\sigma \tag{6.1}$$

Das würde im Falle des medizinischen Textes zu einem Schwellenwert von 140,6, beim linguistischen von 190,2 führen. Das liegt in beiden Fällen weit über dem optimalen Schwellenwert von 22,5 bzw. 55 (siehe Tabelle 4.3). Allein *eine* Standardabweichung liegt in beiden Fällen bereits über dem optimalen Schwellenwert. Es scheint also problematisch, die Standardabweichung in eine Formel für den Schwellenwert miteinzubeziehen.

Im Folgenden möchte ich kurz auf einige Ideen eingehen, die mir als mögliche Lösungen obiger Probleme einfielen. Ob diese Ideen realisierbar sind, bedürfte einer intensiven Prüfung mit einer Vielzahl von Texten, die den Rahmen dieser Arbeit leider bei weitem sprengt.

Fachspezifik

Die Frage, ob es sich bei einem gegebenen Text um eine Einführung oder einen „Expertentext“ handelt, läßt sich vermutlich anhand der durchschnittlichen Werte der statistischen Prüfgröße ermitteln. Wie in Tabelle 6.1 gesehen, hat der medizinische Text beispielsweise hier einen höheren Mittelwert (nämlich 19,2) als der linguistische (15,0). Der Mittelwert ist also so etwas wie ein Maß für die Fachspezifik des Textes.

Das ist natürlich dieselbe Idee, die auch hinter Cohens Ansatz steckt, nur daß ich vorschlage, die Standardabweichung hierbei aus dem Spiel zu lassen.

Die Frage ist, welche Schlüsse man daraus zieht: bei einer hohen Fachspezifik den Schwellenwert der Prüfgröße größer zu wählen als bei niedriger Fachspezifik, ist vermutlich meistens richtig. Beim medizinischen Text hatten wir jedoch gesehen, daß sich bei Verwendung der LR auch mit niedrigen Schwellenwerten gute Ergebnisse erzielen ließen.

Um diese Frage zu klären, müßte man eine große Anzahl von Texten untersuchen und quantitativ analysieren, was hier leider nicht geschehen kann.

Facettenreichtum

Um den Facettenreichtum eines Textes zu berechnen, ließe sich an so etwas wie einen „Clusterwert“ für Terme denken: wie bereits gesagt clustern dieje-

nigen Terme, die nur für ein Teilthema eines Textes wichtig sind, innerhalb relativ kleiner Abschnitte (z.B. Kapitel).

Teilt man nun den Text (automatisch) in seine logischen Bestandteile, d.h. Absätze oder Kapitel, ein, so läßt sich — analog zum Ansatz von [Bookstein 1974] aus Kapitel 2 — der Clusterwert x eines Wortes w definieren als:

$$x(w) = \sum_j \binom{j}{2} s_j(w) = s_2 + 3s_3 + 6s_4 + ... \tag{6.2}$$

s_j ist diesmal die Anzahl von Abschnitten, in denen w genau j-mal auftritt. Man könnte auch die IDF aller Terme berechnen (siehe Ansatz von [Salton 1975]) oder irgendeine andere Maßzahl die groß wird, wenn ein Wort an wenigen Stellen eines Textes gehäuft auftritt.

Sodann ließe sich der Clusterwert des gesamten Textes berechnen als der Mittelwert aller Clusterwerte für Wörter. Ist dieser Clusterwert groß, so handelt es sich offensichtlich um einen Text, der viele verschiedene Themen behandelt.

Zu klären wäre noch die Frage, ab wann ein Clusterwert „groß“ ist bzw. wie sich die Größe des Clusterwertes konkret in der Anpassung der Parameter niederschlagen soll.

Ein vergleichsweise großer Clusterwert sollte aber vermutlich bewirken, daß die Mindestfrequenz für Terme gesenkt wird (wie dies z.B. beim medizinischen Text notwendig ist, zumindest wenn man auch die „Teilbereichsterme“ extrahieren möchte).

Domänenspezifische Morphologie

Einen ähnlichen Ansatz wie für die Fachspezifik halte ich auch bei der Frage nach der Morphologie einer Domäne für möglich: der Mittelwert der statistischen Prüfgröße aller Buchstabentrigramme eines Textes sollte Auskunft darüber geben, wieviel abweichende (also z.B. lateinische oder griechische) Morphologie die gerade betrachtete Domäne aufweist.

Ist der Mittelwert sehr klein, so kann man davon ausgehen, daß sich die Trigrammanalyse kaum lohnt und sich umso mehr auf die Extraktion von Phrasen konzentrieren.

Textlänge

Sicherlich ist der Vorschlag von [Justeson 1995] der einzig gangbare Weg, um dem Parameter Textlänge gerecht zu werden: ist ein Text sehr lang, so müssen höhere Mindestfrequenzen verwendet werden als bei einem kurzen Text. Zu bedenken sind hierbei jedoch zwei Probleme:

1. Es ist ein Proportionalitätsfaktor anzugeben, welcher die Umrechnung Textlänge → Mindestfrequenz ermöglicht. Dieser muß aus der Analyse möglichst vieler Texte gewonnen werden.

2. Die erhaltene Mindestfrequenz muß mit den Ergebnissen des Clusterwertes (s.o.) abgeglichen werden: ein langer Text mit vielen Teilthemen kann eine kleinere Mindestfrequenz erforderlich machen als ein kurzer monothematischer (siehe medizinischer bzw. linguistischer Text)

Eigennamen

Ob ein Text Eigennamen nur in Form von Zitaten oder aber eher als Firmen- oder Produktnamen enthält, läßt sich mit Hilfe flacher, musterbasierter Methoden vermutlich relativ leicht entscheiden: ein Zitat kann nur in einigen wenigen Formaten auftreten; auch Literaturverzeichnisse haben eine charakteristische Gestalt, die sich erkennen läßt.

Vermutlich ist es hier aber doch am einfachsten und am Ende auch sichersten, den Anwender entscheiden zu lassen, ob er Eigennamen mit extrahiert haben möchte oder nicht. Das kostet diesen einen Knopfdruck und das Resultat entspricht dann sicher seinen Wünschen.

Die hier aufgeführten Lösungsansätze zeigen also eine Richtung an, in der sich weiterarbeiten ließe. Ob und wie gut die einzelnen Vorschläge sich in die Praxis umsetzen lassen, muß aber noch geprüft werden.

Im Folgenden soll auf einige mögliche funktionelle Erweiterungen eingegangen werden, mit denen man ein Extraktionssystem noch ausstatten könnte, um seine Ergebnisse besser bzw. in breiteren Kontexten verarbeiten zu können.

6.2 Unterstützung der Terminologiearbeit

Zu Beginn des Kapitels 1 hatte ich kurz erwähnt, daß in der Allgemeinen Terminologielehre die Festlegung von normierten Terminologien angestrebt wird, in denen eine eindeutige Beziehung zwischen Begriffen und Benennungen hergestellt wird.

Dazu muß zu jedem Fachterminus eine Definition existieren, die dabei hilft, ihn klar von anderen Termen abzugrenzen und die eine klare Vorstellung davon gibt, welcher Begriff sich hinter der Benennung verbirgt.

Zusätzlich werden Terminologien meist in hierarchischer Form semantisch angeordnet, d.h. die Termini werden derart zueinander in Beziehung gesetzt, daß Begriffstaxonomien entstehen.

In diesem Abschnitt soll kurz der Frage nachgegangen werden, ob sich der eben beschriebene Vorgang der Terminologiearbeit automatisch unterstützen läßt.

6.2.1 Definitionen

Bei der Erstellung einer genormten Terminologie (z.B. in Form eines Wörterbuches) ist es für einen Terminologen sehr interessant, in Fachtexten enthaltene Definitionen von Fachbegriffen mit diesen zusammen zu extrahieren. In vielen Fällen läßt sich die aus dem Text stammende Definition direkt in das Fachwörterbuch übernehmen.

Die automatische Erkennung von Definitionen kann zunächst über einen einfachen musterbasierten Ansatz versucht werden. Durchforstet man z.B. den linguistischen Text aus Kapitel 4 nach Vorkommen des Musters

„<Term> is a _“

d.h. nach einem Fachterminus, gefolgt von den Wörtern „is a“ und beliebigen weiteren Wörtern, so stößt man auf folgende zwei Sätze (die Termini sind jeweils *hervorgehoben*)

> A *symbol* is a sign in which the relation between form and meaning is arbitrary, based neither upon resemblance or any other natural physical connection.

> *Index* – is a sign used in direct spatial and temporal connection with its meaning, often in the sense of event and consequence: smoke–fire, wind vane–direction of wind.

Sicherlich muß man weitere Muster finden, um die Ausbeute zu erhöhen; die Qualität der gefundenen Definitionen ist aber auf den ersten Blick zufriedenstellend.

Ein weiterer interessanter Aspekt ist das Auffinden von Termen zu einer gegebenen Definition. Dies spielt vielleicht bei der Terminologiearbeit weniger eine Rolle, kann jedoch für einen Laien sehr hilfreich sein, der versucht sich in ein Gebiet einzuarbeiten: er kennt vielleicht den Begriff (d.h. er hat eine Vorstellung davon, um welches Konzept es sich dreht), nicht aber die korrekte Benennung.

Hier kann man mit dem gleichen musterbasierten Kontext ansetzen wie bei der umgekehrten Fragestellung. Zusätzlich wäre aber noch denkbar, die vom Laien eingegebene Definition zu parsen und nach Kollokationen der darin vorkommenden Wörter zu suchen, um die Definition zu finden.

Ein Beispiel: der Laie sucht nach „sign with arbitratry relation between form and meaning“, kennt aber nicht den korrekten Fachbegriff „symbol“. Nun berechnet man die Kollokationen zu jedem der Inhaltswörter, die in der Beschreibung des Laien auftreten: „sign“, „arbitrary“, „relation“, „form“ und „meaning“. Sodann schneidet man diese Mengen und hofft, daß die Schnittmenge nur noch ein Wort (oder zumindest wenige Wörter) enthält. Dieses ist dann vielleicht der gesuchte Fachterm. Ob und wie gut das funktioniert, müßte aber noch untersucht werden.

6.2.2 Begriffsnetze

Der nächste Schritt bei der Terminologiearbeit ist die systematische Anordnung der gefundenen und definierten Termini nach semantischen Kriterien. Meist geschieht dies hierarchisch, also in Form von Taxonomien.

Eine automatische Unterstützung dieser Anordnung müßte in der Lage sein zu erkennen, ob ein Term Kohyponym oder Hyperonym eines anderen ist.

Eine Idee, wie sich zu einem gegebenen Term andere finden lassen, die zu

diesem in paradigmatischer Beziehung stehen, hatte ich in Abschnitt 1.4.3 entwickelt. Dabei wurde nach Termen gesucht, die ähnliche Kollokationen aufweisen wie der Ausgangsterm.

In [Bordag 2003] wird erläutert, wie man zusätzlich Kohyponymie von Hyperonomie unterscheiden kann: die Kollokationssignifikanz zwischen zwei Kohyponymen ist höher als die zwischen Hyponym und Hyperonym.

Ich hatte bereits festgestellt, daß bei den kleinen Texten, mit denen ich hier experimentierte, das Berechnen von statistischen paradigmatischen Beziehungen weitgehend scheitert. Mit größeren Korpora sind aber bestimmt bessere Ergebnisse zu erzielen.

Im Deutschen hilft oft eine einfache Heuristik: Zwei Komposita, die den gleichen Kopf haben, sind meist Kohyponyme, der gemeinsame Kopf ist oft ein Hyperonym (Beispiel: „Leichtöl“ und „Schweröl“ sind Kohyponyme, deren Kopf „Öl“ ein Hyperonym).

Eine hohe Qualität kann aber in naher Zukunft von keinem bisher bekannten Verfahren erwartet werden, wenn es um die automatische Erkennung bestimmter semantischer Relationen wie Hyperonomie oder Kohyponymie geht. Die Anordnung von Termen in Taxonomien bedeutet also weiterhin viel Handarbeit.

6.3 Zweisprachige Terminologie-Extraktion

In vielen Bereichen werden auch Übersetzungen von Fachtermini in verschiedenen Sprachen benötigt. Dies ist z.B. bei der Erstellung von Übersetzungswörterbüchern für bestimmte Fachdomänen (z.B. „Englischwörterbuch für Mediziner“ o.ä.) oder bei der maschinellen Übersetzung (siehe LOGOS) der Fall.

Die zweisprachige Extraktion von Fachterminologie aus sogenannten parallelen Texten (zwei Texte heißen parallel, wenn der eine die Übersetzung des anderen ist) stellt ein eigenes Forschungsgebiet dar.

Die prinzipielle Vorgehensweise beruht hierbei stets auf einer sogenannten *Alignierung*, d.h. der Zuordnung der Wörter oder Sätze eines Quelltextes zu den jeweiligen Übersetzungen im Zieltext. Die meisten Verfahren nehmen erst eine Satzalignierung und — darauf aufbauend — eine Wortalignierung

vor. Sowohl für die Satz- als auch für die Wortalignierung gibt es eine Reihe von Verfahren, die z.B. in [Manning 2002] (S.466 ff.) beschrieben werden.

Um nun bilinguale Terminologie zu extrahieren, führt man zunächst für einen Quelltext eine „normale“ monolinguale Extraktion durch, wie sie in dieser Arbeit beschrieben wurde. Danach versucht man, in dessen Übersetzung (d.h. im Zieltext) Übersetzungskandidaten für jeden Fachbegriff zu finden, indem man Quell- und Zieltext auf Satz- und Wortebene miteinander aligniert.

Zu jedem Fachterminus des Quelltextes erhält man somit eine Reihe von Kandidaten, die meist nach einem Assoziationsmaß (man kann z.B. wieder eine likelihood ratio verwenden) geordnet sind. Nun kann man entweder den besten Kandidaten direkt als Übersetzung akzeptieren oder man läßt den Anwender die richtige Übersetzung auswählen.

Entsprechende Verfahren sind z.B. in [Daille 1994] oder [Dagan 1995] beschrieben. Zu beachten ist dabei noch die Schwierigkeit, Mehrwortbegriffe miteinander zu alignieren, da es z.B. zwischen Deutsch und Englisch sehr oft vorkommt, daß ein englischer Mehrwortbegriff mit einem deutschen Kompositum (also einem Einwortbegriff) übersetzt wird.

Insgesamt sind also noch sehr viele Verbesserungen und Erweiterungen denkbar. Das von mir implementierte System stellt aber eine gute Grundlage für die beschriebenen Erweiterungen dar und erlaubt es — aufgrund der Interaktion mit dem Anwender — Terminologien von hoher Qualität zu produzieren, die dann beliebig weiter verarbeitet werden können.

Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit einem Teilgebiet des Text Mining, versucht also Information (in diesem Fall Fachterminologie) aus natürlichsprachlichem Text zu extrahieren.

Die der Arbeit zugrundeliegende These besagt, daß in vielen Gebieten des Text Mining die Kombination verschiedener Methoden sinnvoll sein kann, um dem Facettenreichtum natürlicher Sprache gerecht zu werden.

Die bei der Terminologie-Extraktion angewandten Methoden sind statistischer und linguistischer (bzw. musterbasierter) Natur. Um sie herzuleiten, wurden einige Eigenschaften von Fachtermini herausgearbeitet, die für deren Extraktion relevant sind. So läßt sich z.B. die Tatsache, daß viele Fachbegriffe Nominalphrasen einer bestimmten Form sind, direkt für eine Suche nach gewissen POS-Mustern ausnützen, die Verteilung von Termen in Fachtexten führte zu einem statistischen Ansatz — der Differenzanalyse.

Zusammen mit einigen weiteren wurden diese Ansätze in ein Verfahren integriert, welches in der Lage ist, aus dem Feedback eines Anwenders zu lernen und in mehreren Schritten die Suche nach Terminologie zu verfeinern. Dabei wurden mehrere Parameter des Verfahrens veränderlich belassen, d.h. der Anwender kann sie beliebig anpassen.

Bei der Untersuchung der Ergebnisse anhand von zwei Fachtexten aus unterschiedlichen Domänen wurde deutlich, daß sich zwar die verschiedenen Verfahren gut ergänzen, daß aber die optimalen Werte der veränderbaren Parameter, ja selbst die Auswahl der angewendeten Verfahren text- und domänenabhängig sind.

Dies zeigt auch die Grenzen des vorgestellten Ansatzes, sowie vieler Text Mining-Verfahren insgesamt auf: der Facettenreichtum der Sprache macht es

auch bei der Kombination mehrerer Verfahren bisher unmöglich, ein System zu konstruieren, welches für alle Texte gleich gut funktioniert.

Die Frage, ob sich dies durch eine „Domänenerkennung“ und anschließende dynamische Anpassung von Parametern doch noch erreichen läßt, konnte in dieser Arbeit nicht mehr beantwortet werden und soll Gegenstand weiterer Forschung sein.

Literaturverzeichnis

[Arppe 1995] Arppe, A. (1995): Term extraction from unrestricted text. NODALIDA-95, Helsinki. URL: http://www.lingsoft.fi/doc/nptool/term-extraction.html, 13.10.2003.

[Biemann 2003] Biemann, C., Bordag, S., Quasthoff, U. (2003): Lernen paradigmatischer Relationen auf iterierten Kollokationen. In: *Proceedings of GermaNet-Workshop: Anwendungen des deutschen Wortnetzes in Theorie und Praxis, Tübingen 2003*, S. 87-94.

[Biemann 2004] Biemann, C., Bordag, S., Heyer, G., Quasthoff, U., Wolff, C. (2004): Language-independent Methods for Compiling Monolingual Lexical Data. In: *Proceedings of CICLing-2004*, S. 214-225.

[Blank 1997] Blank, I. (1997): *Computerlinguistische Analyse mehrsprachiger Fachtexte*. Dissertation, Universität München.

[Bookstein 1974] Bookstein, A., Swanson, D.R. (1974): Probabilistic Models for Automatic Indexing. In: *Journal of the American Society for Information Science*, 25(5), S. 312-318.

[Bordag 2003] Bordag, S., Heyer, G. (2003): Warum ist es möglich, semantische Relationen automatisch zu berechnen? Einige Anmerkungen zur Relevanz strukturalistischer Grundlagen der Semantik. In: *Journal of Quantitative Linguistics* [to appear].

[Bourigault 1992] Bourigault, D. (1992): Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. In: *Proceedings of Coling92*, S. 977-981.

[Brants 2000] Brants, T. (2000): TnT - A Statistical Part-of-Speech Tagger. In: *Proceedings of the 6th Applied Natural Language Processing Conference*, S. 224-231.

[Brill 1992] Brill, E. (1992): A simple rule-based part-of-speech tagger. In: *Proceedings of ANLP-92*, S. 152-155.

[Bußmann 2002] Bußmann, H., Hrsg. (2002): *Lexikon der Sprachwissenschaft.* Stuttgart: Kröner.

[Cohen 1995] Cohen, J.D. (1995): Highlights: language and domain independent automatic indexing terms for abstracting. In: *Journal of the American Society for Information Science*, 46(3), S. 162-174.

[Dagan 1995] Dagan, I., Church, K. (1995): Termight: identifying and translating technical terminology. In: *Proceedings of EACL'95*, S. 34-40.

[Daille 1994] Daille, B., Gaussier, E., Langé, J. (1994): Towards Automatic Extraction of Monolingual and Bilingual Terminology. In: *Proceedings of COLING94*, S. 515-521.

[Damerau 1993] Damerau, F.J. (1993): Evaluating domain-oriented multiword terms from texts. In: *Information Processing and Management*, 29(4), S. 433-447.

[Dillon 1983] Dillon, M., Gray, A. (1983): FASIT: A Fully Automatic Syntactically Based Indexing System. In: *Journal of the American Society for Information Science*, 34(2), S. 99-108

[DIN 2342] Schmalenbach, K. (2002): *DIT - Begriffe der Terminologie.* URL: http://www.dit-online.com/termi/main.htm, 13.10.2003.

[Dunning 1993] Dunning, T. (1993): Accurate Methods for the Statistics of Surprise and Coincidence. In: *Computational Linguistics*, 19(1), S. 61-74

[Evans 1995] Evans, D. A., Lefferts, R. G. (1995): CLARIT-TREC Experiments. In: *Information Processing and Management*, 31(3), S. 385-395.

[Fluck 1985] Fluck, H.-R. (1985): *Fachsprachen, Einführung und Bibliographie.* Tübingen: Francke.

[Frantzi 1996] Frantzi, K. T., Ananiadou, S. (1996): Extracting nested collocations. In: *Proceedings of COLING96*, S. 41-46.

[Heid 1998] Heid, U. (1998): A linguistic bootstrapping approach to the extraction of term candidates from German text. In *Terminology*, 5(2), S. 161-181.

[Heyer 2001] Heyer, G., Läuter, M., Quasthoff, U., Wittig, Th., Wolff, Chr. (2001): Learning Relations using Collocations In: *Proceedings of the IJCAI Workshop on Ontology Learning*, S. 19-24.

[Heyer 2002] Heyer, G., Quasthoff, U., Wolff, C. (2002): Möglichkeiten und Verfahren zur automatischen Gewinnung von Fachbegriffen aus Texten. In: Bullinger, H.-J. (ed.): *Content Management - Digitale Inhalte als Bausteine einer vernetzten Welt*, Fraunhofer IRBN Verlag:Stuttgart, S. 43-49.

[Hoffmann 1988] Hoffmann, L. (1988): *Vom Fachwort zum Fachtext: Beiträge zur angewandten Linguistik.* Tübingen: Gunter Narr Verlag.

[Justeson 1995] Justeson, J.S., Katz, S.M. (1995): Technical terminology: some linguistic properties and an algorithm for identification in text. In: *Natural Language Engineering*, 1(1), S. 9-27.

[Kageura 1996] Kageura, K., Uminuo, B. (1996): Methods of Automatic Term Recognition. A Review. In: *Terminology*, 3(2), S. 259-289.

[Manning 2002] Manning, C.D., Schütze, H. (2002): *Foundations of Statistical Natural Language Processing.* Cambridge, Massachusetts: The MIT Press.

[Maynard 1999] Maynard, D. (1999): *Term Recognition Using Combined Knowledge Sources.* Phd. Thesis, Manchester Metropolitan University.

[Porter 1980] Porter, M. F. (1980): An algorithm for suffix stripping. In: *Program*, 14(3), S. 130-137.

[Quasthoff 2002] Quasthoff, U.; Wolff, C. (2002): The Poisson Collocation Measure and its Applications. In: *Proceedings of the Second International Workshop on Computational Approaches to Collocations*, Wien, Juli 2002 [to appear].

[Riloff 1994] Riloff, E., Lehnert, W. (1994): Information extraction as a basis for high-precision text classification. In: *ACM Transactions on Information Systems*, 12(3), S. 296-333.

[Salton 1975] Salton, G., Wong, A., Yang, C.S. (1975): A Vector Space Model for Automatic Indexing. In: *Communications of the ACM*, 18(11), S.613-620.

[Salton 1988] Salton, G.(1988): Syntactic Approaches to Automatic Book Indexing. In: *Proc. of the 26th Annual Meeting of the Association for Computational Linguistics*, S. 204-210.

[Salton 1989] Salton, G. (1989): *Automatic Text Processing.* Reading, Mass.: Addison-Wesley.

[Salton 1990] Salton, G., Buckley, C. (1990): Improving Retrieval Performance by Relevance Feedback In: *Journal of the American Society for Information Science*, 41(4), S. 288-297.

[Saussure 1967] Saussure, F. de (1967): *Grundfragen der allgemeinen Sprachwissenschaft.* Berlin: DeGruyter.

[Wittgenstein 1978] Wittgenstein, L. (1978): *Philosophical Investigations.*, 3rd edition. Oxford: Blackwell.

[Wüster 1991] Wüster, E. (1991): *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie.* Bonn: Romanistischer Verlag.

Anhang A

Tabellen

Pattern	Recall	Cumulative Recall	Precision
A N	27%	27%	29%
N N	17%	44%	40%
N	16%	60%	16%
N P N	11%	71%	29%
A N N	8%	79%	42%
N P A N	5%	84%	31%
N P N N	4%	88%	33%
A N P N	3%	91%	24%
I N	1%	92%	15%
A A N	1%	93%	21%
N P N P N	1%	94%	26%
N N N	1%	95%	8%
Others	5%	100%	-

Tabelle A.1: Untersuchungen zu Precision und Recall von POS-Filtern aus [Arppe 1995]: Bewertung von 558 gefundenen Termen aus einem Text von 12225 Wörtern

Begriff	Denkeinheit, die aus einer Menge von Gegenständen unter Ermittlung der diesen Gegenständen gemeinsamen Eigenschaften mittels Abstraktion gebildet wird. Anmerkung: Begriffe sind nicht an einzelne Sprachen gebunden, sie sind jedoch von dem jeweiligen gesellschaftlichen und kulturellen Hintergrund einer Sprachgemeinschaft beeinflußt. DIN 2342
Benennung	Aus einem Wort oder mehreren Wörtern bestehende Bezeichnung. Anmerkung 1: Begriffe werden sprachlich durch Benennungen (und Definitionen) repräsentiert. Anmerkung 2: Man unterscheidet zwischen Einwortbenennungen (einschließlich der zusammengesetzten Benennungen) und Mehrwortbenennungen. Kriterium ist die Trennung der Benennungsteile durch Leerstellen. DIN 2342
Bezeichnung	Repräsentation eines Begriffs mit sprachlichen oder anderen Mitteln. DIN 2342
Fachsprache	Bereich der Sprache, der auf eindeutige und widerspruchsfreie Kommunikation in einem Fachgebiet gerichtet ist und dessen Funktionieren durch eine festgelegte Terminologie entscheidend unterstützt wird. DIN 2342
Gemeinsprache	Kernbereich der Sprache, an dem alle Mitglieder einer Sprachgemeinschaft teilhaben. DIN 2342
Terminologie (Fachwortschatz)	Gesamtbestand der Begriffe und ihrer Benennungen in einem Fachgebiet. DIN 2342
Terminologielehre	Wissenschaft von den Begriffen und ihren Benennungen im Bereich der Fachsprachen. DIN 2342
Terminus	Das zusammengehörige Paar aus einem Begriff und seiner Benennung als Element einer Terminologie. DIN 2342

Tabelle A.2: Begriffe der Terminologie (zusammengestellt von/compiled by K. Schmalenbach)

Anhang B

Abkürzungen

Abkürzung	kompletter Begriff
ACT	affix compression trie
HQ	Häufigkeitsquotient
IDF	inverse document frequency
IR	Information Retrieval
KF	Kanonische Form
LR	Likelihood Ratio
MI	Mutual Information
NP	Nominalphrase
POS	part of speech (=Wortart)
TF	term frequency

Anhang C

Begriffe

Begriff	Erklärung
Antonym	Gegenteil
Domäne	wird hier gebraucht im Sinne von: Fachgebiet
Hyperonym	Oberbegriff
Hyponym	Unterbegriff
Kohyponyme	zwei Begriffe sind Kohyponyme, wenn sie einen gemeinsamen Oberbegriff haben.

Erklärung

Ich versichere, daß ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe

Leipzig, den 20. März 2004 Unterschrift