Das Indexieren von natürlichsprachlichen Dokumenten und die inverse Seitenhäufigkeit

Lizentiatsarbeit der Philosphischen Fakultät der Universität Zürich

Referent: Prof. Dr. Michael Hess,
Institut für Computerlinguistik, Universität Zürich

Hollerith 215
honomym 166
Host 282, 493
HTML 311, 366, 413
HTTP 311
Hub 725
Hypermedia 355
Hypertext 355, 1011
IEC 904
Ikone 692
IM s. Informationsmanagement
IMPACT 872
Index s. Schlagwortregister
Indexieren 120
- automatisches 128, 487
- Freitext 130
- gleichordnendes 121, 134
- linguistische Verfahren 130
- Qualität 136
- Vokabular
Indexierungssprache 125
Indexierungstiefe 137
Indikatives Referat 95
Recall 138
Recherche s. Retrieval
Recherchedienst 275
Rechnerverbund 344
Recht s. Informationsrecht
Rechtsnorm 538
Referat 88, 888
Referateorgan 29, 269
Referenzdatenbank 284, 477
Referieren 88
- automatisches 106
- Regelwerk 99
Regeln für die Katalogisierung 65
Regelwerk
- Fernsehen 599
- formale Analyse 64
- Referieren 99
- Rundfunk 508, 593
Register s. Index
Registratur 270
Rehabilitation 741
Relation
- Äquivalenz 171
Schallplatte 508
Schlagwort 181
Schlagwortregister s. Register
Schlitzlochkarte 212
Schriftgutverwaltung 460
Scope note 176
Script 196
Search Engine 313
Sehbehinderter 730
Semantische Beziehung 182
Semantische Zerlegung 168
Semantisches Netz 183
Sender-Empfänger-Modell 16
Serendipity 361
SGML 413, 548, 910, 1011
Sichtlochkarte 213
Software 644
- Betriebssysteme 641
- Datenbanken 649, 1016
- Ergonomie 688
- Thesaurus 179, 657
- Übersetzung 247
Sozialforschung 797

Verfasserin:

Esther Kaufmann
Schwyzerstrasse 64c
CH-8832 Wollerau
Telefon: 01-785 08 70
Email: e.kaufmann@access.unizh.ch

Oktober 2001

Abstract

Die Lizentiatsarbeit gibt im ersten theoretischen Teil einen Überblick über das Indexieren von Dokumenten. Sie zeigt die verschiedenen Typen von Indexen sowie die wichtigsten Aspekte bezüglich einer Indexsprache auf. Diverse manuelle und automatische Indexierungsverfahren werden präsentiert. Spezielle Aufmerksamkeit innerhalb des ersten Teils gilt den Schlagwortregistern, deren charakteristische Merkmale und Eigenheiten erörtert werden. Zusätzlich werden die gängigen Kriterien zur Bewertung von Indexen sowie die Masse zur Evaluation von Indexierungsverfahren und Indexierungsergebnissen vorgestellt. Im zweiten Teil der Arbeit werden fünf reale Bücher einer statistischen Untersuchung unterzogen. Zum einen werden die lexikalischen und syntaktischen Bestandteile der fünf Buchregister ermittelt, um den Inhalt von Schlagwortregistern zu erschliessen. Andererseits werden aus den Textausschnitten der Bücher Indexterme maschinell extrahiert und mit den Schlagworteinträgen in den Buchregistern verglichen. Das Hauptziel der Untersuchungen besteht darin, eine Indexierungsmethode, die auf linguistikorientierter Extraktion der Indexterme und Termhäufigkeitsgewichtung basiert, im Hinblick auf ihren Gebrauchswert für eine automatische Indexierung zu testen. Die Gewichtungsmethode ist die inverse Seitenhäufigkeit, eine Methode, welche von der inversen Dokumentfrequenz abgeleitet wurde, zur automatischen Erstellung von Schlagwortregistern für deutschsprachige Texte. Die Prüfung der Methode im statistischen Teil führte nicht zu zufriedenstellenden Resultaten.

Inhaltsverzeichnis

1 Einführung 1
1.1 Problemstellung 1
1.2 Aufbau der Arbeit 4

2 Das Indexieren von Dokumenten 6
2.1 Der Index und seine Funktionen 7
2.2 Ebenen des Indexierens 8
2.3 Indextypen 9
2.4 Die Indexsprache 18
2.4.1 Der Indexterm 19
2.4.2 Der Indexsprachenwortschatz 20
2.4.2.1 Begriffsdefinitionen 20
2.4.2.2 Kontrolliertes Vokabular 22
2.4.2.2.1 Thesaurus 23
2.4.2.2.2 Klassifikation 27
2.4.2.2.3 Schlagwortliste 30
2.4.2.2.4 Classaurus 30
2.4.2.3 Freies Vokabular 31
2.4.2.4 Die Anforderungen an ein Indexierungsvokabular 32
2.4.3 Die Indexsprachengrammatik 33
2.4.3.1 Die Segmentbildung 35
2.4.3.2 Die Schlagwortkette 36
2.4.3.3 Der Relationenindikator 36
2.4.3.4 Links und Proximity-Operatoren 37
2.4.3.5 Topologische Verfahren 38
2.5 Indexierungsverfahren 39
2.5.1 Manuelle Verfahren 40
2.5.1.1 Extraktionsmethoden 40
2.5.1.1.1 Manuelle Stichwortindexierung 41
2.5.1.2 Zuteilungsmethoden 41
2.5.1.2.1 Freies Indexieren 42
2.5.1.2.2 Kontrolliertes Indexieren 43
2.5.1.2.3 Klassifizierendes Indexieren 44
2.5.1.2.4 Verbindliches Indexieren 45
2.5.1.2.5 Hybrid-Indexieren (manuell zuteilend) 46

2.5.2 Automatische Verfahren ... 46
2.5.2.1 Extraktionsmethoden ... 48
2.5.2.1.1 Volltextspeicherung ... 48
2.5.2.1.2 Linguistische Verfahren ... 49
2.5.2.1.3 Statistische Verfahren ... 54
2.5.2.1.3.1 Termgewichtung ... 54
2.5.2.1.3.2 Vektorraummodelle ... 58
2.5.2.1.3.3 Probabilistische Retrievalmodelle ... 58
2.5.2.1.4 Andere Indexierungsverfahren ... 59
2.5.2.1.5 Hybrid-Indexierung (automatisch) ... 59
2.5.2.2 Zuteilungsmethoden ... 60
2.5.2.2.1 Thesaurusbasierte Verfahren ... 61
2.5.2.2.2 Klassifizierende Verfahren ... 61
2.5.2.3 Probleme der automatischen Indexierung ... 63
2.5.3 Die Indexierungsverfahren im Vergleich ... 64

3 Das Schlagwortregister ... 68
3.1 Das Schlagwortregister und seine Funktionen ... 68
3.2 Die Bestandteile eines Schlagwortregisters ... 70
3.2.1 Die Einträge ... 70
3.2.1.1 Das Schlagwort ... 71
3.2.1.2 Die Untereinträge ... 72
3.2.1.3 Die Fundstellen ... 73
3.2.1.4 Die Querverweise ... 74
3.3 Schlagwortregistertypen ... 75
3.3.1 Namen- und Sachregister ... 75
3.3.2 Das alphabetische Schlagwortregister ... 76
3.3.3 Das systematische Schlagwortregister ... 77
3.4 Das manuelle Vorgehen beim Erstellen eines Schlagwortregisters ... 79
3.4.1 Inhaltsanalyse ... 80
3.4.2 Übersetzung in Indexterme ... 82
3.4.3 Nachbearbeitung ... 83
3.5 Form und Layout eines Schlagwortregisters ... 84
3.5.1 Schlagwörter und Untereinträge ... 86
3.5.2 Fundstellen ... 89
3.5.3 Querverweise ... 90
3.5.4 Anordnung der Einträge ... 91
3.5.5 Layout des Registers ... 92
3.6 Die Anforderungen an ein Schlagwortregister ... 95

4 Evaluation eines Index 99
4.1 Merkmale eines „guten“ Index 100
4.2 Faktoren, welche die Indexierungsqualität beeinflussen 101
4.3 Evaluation mit vier Bewertungskriterien 102
4.4 Evaluation mittels Retrievaltests 104
4.4.1 Fehlerstatistiken 106
4.4.2 Konsistenz 106
4.4.3 Relevanz 107
4.4.4 Recall und Precision 110
4.4.5 Fallout 114
4.4.6 Accuracy 115
4.4.7 Error 115
4.4.8 E-Wert 116
4.4.9 F-Wert 116

5 Statistische Untersuchungen 118
5.1 Material und Methoden 118
5.2 Untersuchung der Schlagwortregister 122
5.2.1 Die Schlagwörter und Untereinträge 122
5.2.2 Die Wortkategorien der Einträge 126
5.2.3 Die Bestandteile der Einträge 137
5.2.4 Zusammenfassung 150
5.3 Indexierung der Textkörper 150
5.3.1 Die inverse Seitenhäufigkeit 152
5.3.2 Ergebnisse 155
5.3.2.1 Anpassung der Bewertungsmasse 156
5.3.2.2 Die einzelnen Textausschnitte 159
5.3.2.3 Alle Textausschnitte 165
5.3.2.3.1 Analyse der gefundenen relevanten Indexterme 169
5.3.2.3.2 Analyse der nicht gefundenen relevanten Indexterme 171
5.3.2.3.3 Analyse der gefundenen nicht relevanten Indexterme 173
5.3.3 Zusammenfassung 175

6 Zusammenfassung und Schlussfolgerungen 176

7 Ausblick 180

8 Literaturverzeichnis 183
8.1 Untersuchte Bücher 183
8.2 Sekundärliteratur 183

Verzeichnis der Tabellen

Tabelle 1: Vergleich der Indexierungsverfahren nach Fugmann 65
Tabelle 2: Richtlinien für Registerlängen nach Mulvany 85
Tabelle 3: Faktoren, welche die Indexierungsqualität beeinflussen können 102
Tabelle 4: Quantitäten der Dokumente in einer Kollektion 111
Tabelle 5: Länge der untersuchten Buchindexe 119
Tabelle 6: Anzahl Registereinträge pro Buchseite 120
Tabelle 7: Anzahl Schlagwörter und Untereinträge in den untersuchten Registern 122
Tabelle 8: Anzahl Tokens in den Schlagwörtern und Untereinträgen 124
Tabelle 9: Tokens pro Schlagwort, Untereintrag und Eintrag 125
Tabelle 10: Lexikalische Wortkategorien in den fünf Registern 127
Tabelle 11: Lexikalische Wortkategorien in Prozent zum Total der Tokens 129
Tabelle 12: Wortkategorien zusammengefasst 1 130
Tabelle 13: Wortkategorien zusammengefasst 2 131
Tabelle 14: Wortkategorieklassen in den Schlagwörtern und Untereinträgen 133
Tabelle 15: Wortkategorieklassen zusammengefasst in den Schlagwörtern und Untereinträgen . 135
Tabelle 16: Substantive: Komposita, Plural- und Singularformen 136
Tabelle 17: Zusammensetzung der Schlagwörter 142
Tabelle 18: Zusammensetzung der Untereinträge 148
Tabelle 19: Anzahl Sätze und Tokens in den Textkörperausschnitten 152
Tabelle 20: Anzahl extrahierte nominale Wortgruppen sowie erzielte tf * ipf-Werte 155
Tabelle 21: Quantitäten der Indexterme in Textausschnitten und dazugehörigen Registern 156
Tabelle 22: Quantitäten der Indexterme bei Cap (w_{min} >= 1,71) 159
Tabelle 23: Erreichte Evaluationswerte bei Cap 160
Tabelle 24: Quantitäten der Indexterme bei Hausser (w_{min} >= 1,71) 161
Tabelle 25: Erreichte Evaluationswerte bei Hausser 161
Tabelle 26: Quantitäten der Indexterme bei Linke et al. (w_{min} >= 1,71) 162
Tabelle 27: Erreichte Evaluationswerte bei Linke et al. 162
Tabelle 28: Quantitäten der Indexterme bei Stegmüller (w_{min} >= 1,71) 163
Tabelle 29: Erreichte Evaluationswerte bei Stegmüller 163
Tabelle 30: Quantitäten der Indexterme bei Werlen (w_{min} >= 1,71) 164
Tabelle 31: Erreichte Evaluationswerte bei Werlen 164
Tabelle 32: Quantitäten der Indexterme bei allen Textausschnitten (w_{min} >= 1,71) 166
Tabelle 33: Erreichte Evaluationswerte bei allen Textausschnitten (w_{min} >= 1,71) 166
Tabelle 34: Quantitäten der Indexterme bei allen Textausschnitten (w_{min} >= 2,01) 167
Tabelle 35: Erreichte Evaluationswerte bei allen Textausschnitten (w_{min} >= 2,01) 168
Tabelle 36: Anzahl der gefundenen relevanten Indexterme und erzielte tf * ipf-Werte 169
Tabelle 37: Bestandteile der gefundenen relevanten Indexterme 170
Tabelle 38: Bestandteile der relevanten nicht gefundenen Indexterme 172
Tabelle 39: Bestandteile der gefundenen nicht relevanten Indexterme 174

1 Einführung

1.1 Problemstellung

Jegliches Arbeiten im Berufs- und Privatleben setzt voraus, dass man sich Wissen aneignet und Neuerungen verfolgt. Informiertsein ist eine Notwendigkeit; Information ist ein bedeutendes Gut und meint jede Nachricht, die für einen Empfänger von Interesse ist. Da es nicht möglich ist, über alle Bereiche, mit welchen man konfrontiert wird, informiert zu sein und sich an alle Informationen, die man einmal gesammelt hat, erinnern kann, müssen zusätzlich zum eigenen Erinnerungs- und Assoziationsvermögen weitere Hilfsmittel zum Finden oder Wiederfinden von Informationen geschaffen werden. Es ist die Aufgabe der **Inhaltserschliessung**, das Aufspüren und die Wiederauffindung von relevanten Informationen zu ermöglichen. Eine Art von Inhaltserschliessung ist das **Indexieren**.

Der Prozess des Indexierens weist einem Originaltext inhaltskennzeichnende **Begriffsbenennungen** zu und kreiert eine kurze Beschreibung, eine Charakterisierung des Originaltextes. Das Resultat ist eine Textrepräsentation oder ein Index in einem bestimmten Format. (Das Gleiche gilt auch für Abstracts oder Kurzreferate.) Die Begriffsbenennungen oder Indexterme können Wörter der natürlichen Sprache sein oder z.B. Notationen aus einer Klassifikation. Die Essenz von Dokumenten wird mit einer Indexierung festgehalten und muss zum Zeitpunkt einer späteren Suche rekonstruierbar sein. Eine Suche nach Informationen vollzieht sich mit einer Fragestellung bestehend aus Suchtermen, die mit den Indextermen der Dokumente oder Texte abgeglichen werden. Erst durch die besondere Erschliessung und Dokumentation von Inhalten durch Indexierung können Informationen im Bedarfsfall leicht und zuverlässig wieder auffindbar und wieder verwendbar gemacht werden, und zwar ohne dass der Mensch hierbei bezüglich seines Suchvermögens überfordert wird und ohne dass beim Indexierungsprozess untragbar grosse Kosten entstehen.

Das Wiederauffinden von Informationen bezeichnet man als **Retrieval**. Beim Retrieval mithilfe von Indextermen handelt es sich um **Referenz-Retrieval**, also diejenige Form der Dokumentation, die auf eine Benutzerfrage nicht die Antwort selbst, sondern (möglichst) genau diejenigen Dokumente liefert, aus denen die Antwort ersehen werden kann.[1] Die vorliegende Arbeit bezieht sich durchwegs auf Referenz-Retrieval. Das Ziel ist nicht die Information selbst, sondern ein Zeiger zur relevanten Information, wobei der Index als indikative Textrepräsentation fungiert.[2]

[1] Knorz 1983:1
[2] Fugmann 1999:146; Moens 2000:61

Indexierungen werden für Recherchen in verschiedenartigen Informationssystemen zu unterschiedlichen Zwecken ausgeführt, z.B. für Katalogsysteme in Bibliotheken, für Information Retrieval-Systeme in grossen Dokumentkollektionen, für Schlagwortverzeichnisse in Büchern, für Datenbanken, für Browsing- oder Navigationssysteme im Internet usw. Ein Information Retrieval-System selektiert beispielsweise Dokumente aus einer Dokumentkollektion als Antwort auf eine **Query** (Benutzeranfrage, Fragestellung, Anfrage, Suchanfrage, Suchauftrag oder Suchbedingung). Der Begriff *Information Retrieval* bezieht sich manchmal auf Informationsmanagement ganz allgemein, öfter jedoch meint er die Wiedergewinnung von Dokumenten, die ein bestimmtes Informationsbedürfnis befriedigen.[3] Bei allen Informationssystemen werden die Terme der Textrepräsentation mit den Termen der Suchanfrage abgeglichen, die eine formale Repräsentation des Benutzerbedürfnisses darstellen. Eine Anfrage besteht aus einem einzigen Term oder einer Kombination von Termen. Als Ergebnis wird eine Liste von Verweisen auf diejenigen Texte, die beim Vergleich der beiden Termarten am besten abgeschnitten haben, an den Benutzer geliefert. Werden hinsichtlich einer Anfrage relevante Texte nicht als Antwort zurückgegeben, so entsteht für den Suchenden ein **Informationsverlust**. Erhält er Texte, die für sein Informationsbedürfnis nicht von Bedeutung sind, muss er sich mit **Ballast** auseinander setzen.

In den letzten Jahren ist die Informationsflut stark angestiegen, und die Menge der produzierten Information wird in absehbarer Zeit weiter anwachen. Zusätzlich werden wir gegenwärtig mit einer riesigen Menge an elektronischen Dokumenten, die in natürlicher Sprache verfasst sind, konfrontiert. Dabei wird es immer wichtiger, die Informationen über die Informationen zu organisieren und zu verwalten bzw. die Informationsinhalte z.B. mittels Textrepräsentationen zu managen. Der Bedarf nach automatischem Indexieren wird entsprechend immer dringender. Allerdings arbeitet die gegenwärtige Indexierungs- und Recherchenpraxis der automatischen Verfahren mit erheblichen Quoten an Informationsverlust und Ballast. Aber auch bei vielen manuell erstellten Buchregistern muss man sich mit hohem Informationsverlust abfinden. Schuld an dieser Sachlage ist der Umstand, dass die einzelnen Dokumente einer Sammlung oder die einzelnen Passagen eines Buches für das Wiederfinden nicht genügend gründlich und nicht genügend sorgfältig indexiert worden sind. Offenkundig wird Informationsverlust oft nur bei mangelhaften Registern zu Büchern, da dort der Überblick über den Inhalt und die Gegenstände des Buches möglich ist, insbesondere wenn ein Benutzer das Buch z.B. schon einmal gelesen hat und etwas nachschlagen will, die einschlägige Textstelle jedoch nicht mehr in Erinnerung hat. Hohe Effektivität bei der Wiedergewinnung von Informationen wird mit relativ grossem Aufwand bei der Indexierung erkauft. Dabei ist dennoch zu betonen, dass ein Retrievalergebnis stets nur das an einen Benutzer liefern kann, wonach der Benutzer auch tatsächlich gesucht hatte.

Der Prozess der Indexierung wird oft als Kunst verstanden. Dies bezeugen diverse Buchtitel aus dem Gebiet des Indexierens (*The Art of Indexing*[4] etc.). Für viele Fachleute liegt in der Indexierung von Dokumenten und Büchern eine Aufgabe hohen Ranges und grossen Umfangs. Im

[3] Moens 2000:26
[4] Bonura 1994; Knight 1979

angloamerikanischen Ausland hat sich die Profession des freiberuflichen Indexierers bereits seit langem fest etabliert. Dort gibt es wissenschaftliche Gesellschaften (z.B. *American Society of Indexers*), in denen das Buchindexieren gelehrt wird und in deren Verbandszeitschriften (z.B. *Key Words* oder *The Indexer*[5]) die beruflichen Aspekte der Ausübung von professioneller Indexierungsarbeit erörtert werden. In der gesamten Indexiererfachwelt scheint sich eine Zweiteilung ergeben zu haben: Auf der einen Seite stehen die professionellen, manuell und computerunterstützt arbeitenden Indexierer (sie indexieren Bücher, Artikel, Fachzeitschriften usw. und sind in einem Verlag oder in Bibliotheken angestellt); auf der anderen Seite befindet sich die Gruppe von Leuten, die sich mit dem Indexieren von riesigen Dokumentkollektionen, vor allem im Bereich des Internet, beschäftigen. Diese verwenden automatische Methoden. Beide Seiten sind der Auffassung, dass ihre Methoden die besten Indexierungs- und Retrievalergebnisse produzieren und beurteilen die Arbeit der Gegenseite eher abwertend. Dass man bei den beiden Lagern von ganz anderen Ansprüchen und Zielen ausgeht und sich die beiden Seiten auch ergänzen können, wird häufig übersehen. Im Hinblick auf die weiterhin anwachsende Menge von natürlichsprachlichen Texten und Dokumenten wird das manuelle Indexieren aus Gründen der Kosten- und Zeiteinsparung vermehrt auf die automatischen Verfahren ausweichen müssen. Die in den nachfolgenden Kapiteln gemachten Ausführungen wollen versuchen, in dieser Richtung einen Beitrag zu leisten.

Es ist das Ziel der Arbeit, eine möglichst umfassende Übersicht über die Arten und Funktionen von Indexen, die Charakteristiken einer Indexsprache und die manuellen sowie automatischen Verfahren und Techniken des Indexierens zu geben. Schwergewichtig sollen die Schlagwortregister, die sich ganz hinten in Sachbüchern finden lassen, hinsichtlich ihrer Eigenschaften und Eigenheiten untersucht wie auch die verschiedenartigen Aspekte einer Evaluation von Indexen aufgezeigt werden. Ferner gilt es, herauszuarbeiten, was in realen Schlagwortverzeichnissen steht, und zu testen, ob sich die Methode der inversen Seitenhäufigkeit für das automatische Erstellen von Schlagwortregistern für deutschsprachige Texte eignet. Ich gehe dabei von der Hypothese aus, dass die komplexen kognitiven Prozesse beim manuellen Indexieren nicht einfach maschinell simuliert werden können, sondern dass auf Techniken und Verfahren aus anderen Gebieten ausgewichen werden muss.

Die Idee für die vorliegende Arbeit entstand durch den Umstand, dass die zu den Vorlesungen der Computerlinguistik an der Universität Zürich vorhandenen schriftlichen Unterlagen oder Vorlesungsskripte zum Teil über 500 A4-Seiten lang sind. Es ist entsprechend schwierig, in einem solchen Skript einen ganz bestimmten Sachverhalt oder eine Begriffsdefinition zu finden, wenn der Suchende mit dem Gebiet der Computerlinguistik nicht so vertraut ist oder wenn ein gesuchter Ausdruck nicht im Inhaltsverzeichnis vorkommt. Es wäre von grosser Nützlichkeit, wenn die Skripte einen Index, ein Namen- und Schlagwortverzeichnis besitzen würden, welche die Recherche nach einzelnen Themen oder Erläuterungen erleichtern würden. Da sich die Inhalte der

[5] American Society of Indexers 2000

Vorlesungsskripte aufgrund von Anpassungen an die aktuellen Verhältnisse verändern können, sollte eine Erstellung eines solchen Index automatisch realisierbar sein. Aus dieser Bedarfslage entstand die Idee, eine Lizentiatsarbeit zum Thema des Indexierens zu schreiben, welche einen Überblick über das Sachgebiet darstellt und eine bestimmte Technik des automatischen Indexierens testet.

1.2 Aufbau der Arbeit

Die Arbeit gliedert sich in zwei Hauptteile: Der erste Teil zeigt die in der Fachwelt gebräuchlichen theoretischen Grundlagen des Indexierens auf und umfasst die Kapitel 2, 3 und 4; der zweite Teil widmet sich im Kapitel 5 den statistischen Untersuchungen zur Ermittlung von Schlagwortregisterinhalten und zur Überprüfung des Gebrauchswertes der inversen Seitenhäufigkeit für die Erstellung von Schlagwortregistern.

Das 2. Kapitel beschäftigt sich allgemein mit dem Indexieren von Dokumenten. Es werden die verschiedenen Ebenen des Indexierens und Typen von Indexen vorgestellt. Ausgehend von den anfänglich festgelegten Begriffsdefinitionen, setzen wir uns ferner mit dem Wortschatz und der Grammatik von Indexsprachen auseinander, wobei die Unterscheidung von freien und kontrollierten Vokabularien mit ihren jeweiligen Vorzügen und Nachteilen fokussiert wird. Danach werden die manuellen sowie automatischen Indexierungsverfahren und -techniken präsentiert. Neben den Zuteilungsmethoden wird insbesondere auf die Extraktionsmethoden des maschinellen Indexierens eingegangen. Es werden die gängigen linguistischen und statistischen Verfahren vorgeführt. Das Kapitel wird mit einem kritischen Vergleich der diversen Verfahren abgeschlossen.

Obschon Kapitel 2 das Indexieren von Dokumenten und Texten generell behandelt, betreffen die darin gemachten Ausführungen auch einen besonderen Typ von Index, das Schlagwortregister. Im 3. Kapitel werden die spezifischen Merkmale und Eigenheiten von Schlagwortregistern dann detailliert betrachtet. Es werden die Bestandteile eines Schlagwortverzeichnisses aufgezeigt, die zwei grundsätzlichen Schlagwortregistertypen vorgestellt und die Form sowie das Layout von Buchregistern erforscht. Daneben beschäftigen wir uns auch mit dem manuellen Erstellen von Schlagwortregistern, um daraus mögliche Erkenntnisse für eine automatische Inhaltserschliessung abzuleiten.

Die Frage nach den charakteristischen Merkmalen von guten und nützlichen Indexen führt zu Kapitel 4. Die wichtigsten Kriterien zur Evaluation von Indexen werden erläutert sowie die Masse zur Performanzmessung von Indexierungssystemen. Der Schwerpunkt liegt auf der Vorstellung der auf dem Kriterium der Relevanz basierenden Bewertungsmasse Recall sowie Precision. Er wurde ausgewählt, weil die Masse bei der Auswertung der statistischen Untersuchung Anwendung finden.

Dann erfolgt der Übergang zum zweiten, statistischen Teil der Arbeit. Das Kapitel 5 stellt als Kernstück die Termgewichtungsmethode der inversen Seitenhäufigkeit vor. Vorab werden die Resultate der Untersuchung von fünf realen Buchregistern vorgeführt. Dies geschieht im Hinblick auf eine Ergründung der lexikalischen und syntaktischen Bestandteile von Schlagwortregistereinträgen. Erst danach unterziehen wir Buchtextausschnitte, die zu den erforschten Registern gehören, einer Analyse und lassen sie mit Hilfe der inversen Seitenhäufigkeit automatisch indexieren. Im Anschluss daran werden die Ergebnisse ausgewertet und Erkenntnisse gezogen.

Das Kapitel 6 stellt die in der Arbeit aufgeführten wichtigsten Punkte aus der Fachwelt des Indexierens sowie die aus den statistischen Untersuchungen resultierenden Ergebnisse nochmals zusammenfassend dar und zieht Schlussfolgerungen.

Ein Ausblick auf die zukünftigen Arbeiten und Verarbeitungsschritte zur Verbesserung des Indexierungsverfahrens der inversen Seitenfrequenz bildet das Kapitel 7. Das Literaturverzeichnis in Kapitel 8 schliesst die Arbeit ab.

Bevor ich nun zum ersten Teil übergehe, möchte ich noch drei organisatorische Bemerkungen anfügen:

- Im Sinne der Lesbarkeit wurde in dieser Arbeit auf eine konsequente Doppelnennung der weiblichen und männlichen Formen zugunsten der männlichen Formen verzichtet.

- Die Orthografie orientiert sich an der amtlichen Regelung der deutschen Rechtschreibung von 1996.

- Bezüglich Typografie und Zeichensetzung ist zu beachten: Zum ersten Mal auftretende Fachausdrücke werden durch Fettdruck markiert. Dort wo es wichtig erscheint, die englische Terminologie anzugeben, stehen die entsprechenden Begriffe in Klammern, sofern sie zur Verfügung standen. Alle englischen Terme sowie Ausdrücke wie in „der Begriff *Schlagwort*“ sind kursiv gedruckt. Im Lauftext auftretende Beispiele oder im übertragenen Sinn gemeinte Ausdrücke werden durch Anführungs- und Schlusszeichen gekennzeichnet.

TEIL 1

2 Das Indexieren von Dokumenten

Schriftlich oder bildlich niedergelegtes Wissen bedarf einer besonderen Erschliessung, um in einem späteren Bedarfsfall wieder auffindbar zu sein. Diese Auffindung kann indirekt über die Anlage eines Index erfolgen. Ein **Index** (*index*) ist also ein methodischer, ordentlicher Führer zum intellektuellen Inhalt und zum physikalischen Ort von Informationen. Die gewünschte Information selbst ist im Index nicht enthalten, er ist vielmehr der **Zeiger** (*pointer*) zu diesen Wissensaufzeichnungen.[6] Der Index selbst besteht aus einer Menge von **Einträgen** (*entries*) oder **Indextermen** (*index terms*), welche die Quelle der Information kennzeichnen. Die Information der Indexterme ist Metainformation[7], die den Benutzer eines Index zu den von einem Autor ausgedrückten Themen und Ideen lenkt.

Mit **Indexieren** wird der Prozess bezeichnet, bei welchem die Einträge eines Index produziert werden. Dieser Prozess ist zweistufig: 1. Inhaltsanalyse zur Erkennung der wieder auffindbar zu machenden Essenz eines Textes, 2. Zuweisung der Indexterme zur Wiedergabe dieser Essenz in einer ausreichend spezifischen, gut voraussehbaren und ausreichend wiedergabetreuen Form.[8] Schliesslich gehört auch die Auswahl der physischen Form des endgültigen Index zum Indexierungsprozess.

Das grundlegende Problem beim Konstruieren eines Index ist die präzise Repräsentation eines Dokuments mit einer überschaubaren Anzahl von Indextermen, wobei das Dokument selbst aus Tausenden von Wörtern bestehen kann. Die Indexterme müssen so gewählt sein, dass die wesentliche Bedeutung und der Inhalt eines Dokuments von diesen Einträgen aus erschlossen werden können. Ein Index wird nicht nur anhand der Zeitersparnis bei der Informationssuche beurteilt, sondern auch aufgrund der Qualität, mit der er die in einem Dokument behandelten Themen reflektiert.

Indexe können sehr unterschiedliche Formen haben. Viele werden publiziert und auf einer nationalen oder internationalen Ebene verbreitet. Andere können hausinterne Indexe sein, um den Zugang zu lokalen Informationsdatenbanken zu gewährleisten. Indexe sind vielleicht in einem Appendix zu einer Zeitschrift oder als Zeitschrift selbst anzutreffen, oder sie sind Teil eines publizierten Klassifikationsschemas. Ein Index kann als Schlagwortregister in einem Buch enthalten sein sowie auf Karten vorliegen, in Mikroform oder in maschinenlesbarer Form. Eines ist jedoch allen Indexen gemeinsam: Sie repräsentieren den Inhalt eines Dokuments auf eine geordnete

[6] Cleveland/Cleveland 1990:2
[7] Lahtinen 2000:29
[8] Fugmann 1999:216

Weise, sodass ein Benutzer so direkt und so schnell wie möglich zur benötigten Information gelangt.

In diesem Kapitel wurde bisher stets von Dokumenten und nicht von Texten gesprochen. Üblicherweise versteht man unter einem **Dokument** irgendeine gedruckte Repräsentation, die textuelle und/oder nichttextuelle Komponenten enthält, mit der Absicht der Wissensvermittlung. Als **Text** wird dagegen eine Gruppe linguistischer Einheiten, die zu einem Konglomerat mit einer kommunikativen Absicht verknüpft sind, bezeichnet.[9] Für den weiteren Verlauf der Arbeit ist eine exakte Unterscheidung nicht wirklich bedeutend; die beiden Begriffe werden generell synonym verwendet. Im Umfeld des Information Retrieval werde ich eher von Dokumenten sprechen, im Zusammenhang mit Schlagwortregistern und der Untersuchung der Buchregister wird überwiegend von Texten die Rede sein, und zwar von expositorischen Texten, also nicht narrativen. Vermehrt wird heute auf die Notwendigkeit hingewiesen, dass sich die Indexierungsindustrie auch mit dem Indexieren von gesprochener Sprache, von Fiktionsliteratur und vor allem von bildlich dargestellter Information auseinander setzen sollte.[10]

2.1 Der Index und seine Funktionen

Die Hauptaufgabe eines Index besteht darin, Repräsentationen von publizierten Elementen zu konstruieren. Dies soll in einer Form geschehen, die das Einschliessen dieser Repräsentationen in irgendeine Art von Datenbank erlaubt.[11] Die Datenbank kann in gedruckter Form, in elektronischer Form oder in Form von Karten (z.B. bei einem traditionellen Katalogsystem in einer Bibliothek) sein. Der Indexierungsprozess identifiziert die in einem Dokument behandelten Themen, während die Dokumente selbst meistens in eine andere Art von Datenbank eingefügt werden, z.B. in die Regale in einer Bibliothek. Die Repräsentationen ermöglichen einen indirekten Zugriff auf den Inhalt eines Dokuments. Im Gegensatz dazu ermöglicht ein Abstract oder Kurzreferat einen direkten Zugriff.[12] Es sollte dabei nicht vergessen werden, dass ein Index auch über den physischen Standort des Dokuments bzw. der Information Auskunft geben muss.

Obschon alle Arten von Indexen die gleiche Funktion haben, so gibt es verschiedene Ebenen von Indexierungen. Innerhalb dieser Ebenen wiederum existieren mehrere Typen von Indexen. Diese Ebenen und Typen werden im Folgenden dargestellt.

[9] Moens 2000:4f.
[10] Cleveland/Cleveland 1990:126; Mulvany 2000:7
[11] Lancaster 1998:1
[12] Knorz 1997:120

2.2 Ebenen des Indexierens

Ein Dokument kann indexiert werden, indem die in einem Text tatsächlich verwendeten Wörter als Basis dienen. Solche **Wort- und Namenindexe** (*word and name indexes*) werden mit den von einem Autor gebrauchten Namen, Bezeichnungen und Wörtern indexiert mit dem Ziel, dass der Index die Informationen und Ideen, die der Autor bei der Kreierung des Textes in seinem Kopf hatte, sehr genau abbildet. Wort- oder Namenindexe werden manchmal auch **Kondordanzen** (*concordances*) genannt. Eine Konkordanz wird definiert als eine alphabetische Liste von wichtigen Wörtern aus einem Text mit Angabe der zugehörigen Fundstellen.[13] Diese Indexe sind wertvoll für Benutzer, z.B. Linguisten, die einen exakten Term oder ein genaues Wort im Kontext des Textes und seine Lokalisation benötigen. Ansonsten sind Konkordanzen in ihrer Nützlichkeit eher beschränkt. Diese Art von Indexierung kann sehr einfach maschinell vollzogen werden und ist somit schnell sowie billig auszuführen. Der grosse Nachteil ist, dass die Suche nach Informationen etwas kompliziert ist. Eine Suche gestaltet sich schwierig und unsicher, da ein Wort- oder Namenindex ähnliche Einträge über sehr viele synonyme Ausdrücke verteilt, ohne dabei Rechtschreibfehler zu berücksichtigen bzw. diese zu ignorieren, und er bringt jede allgemein-spezifisch-Beziehung zwischen Ausdrücken durcheinander, die jedoch in der unbewusst gebrauchten Indexsprache existieren. Konkordanzartige Indexe sind zeitraubend, sie übertragen bei ihrer Benutzung die Last auf den Suchenden. Die Geschwindigkeit und die Ökonomie, die bei der Schaffung eines solchen Index erreicht werden, werden also mit dem Zeitaufwand und der Anstrengung des Benutzers bezahlt. Das häufigste Beispiel eines Originaltextes zur Erstellung von Konkordanzen ist die Bibel.[14]

Ein grosser Teil aller Indexe sind **Buchindexe** (*book indexes, back-of-the-book-indexes*). Tatsächlich ist es so, dass beim Erwähnen des Wortes *Index* die meisten Menschen an einen Buchindex denken.[15] Ein Buchindex besteht aus einer Liste von normalerweise alphabetisch geordneten Wörtern am Ende eines Buches, die zu jedem Wort oder Namen mindestens eine Seitenzahl (manchmal auch eine Paragrafennummer) angibt, wo das entsprechende Thema im Buch behandelt wird. Ein Buchindex lokalisiert die in einem Buch enthaltenen Informationen, damit ein Leser nicht das ganze Buch lesen oder nochmals lesen muss. Es gilt für Buchindexe (ebenso wie für Kondordanzen), dass der Index kein Ersatz für die Informationen im Buch ist, sondern lediglich ein Verweis auf die Informationen. Den Buchindexen oder Schlagwortregistern ist in dieser Arbeit ein eigenes Kapitel gewidmet (Kapitel 3).

Periodische Indexe (*periodical indexes*) repräsentieren den Inhalt von Zeitschriften. Sie unterscheiden sich von Buchindexen durch ihren breiteren Themenbereich, was zu Problemen führen kann. Während das Erstellen eines Buchindexes einen klaren Anfang und Schluss aufweist und von einer einzigen Person ausgeführt werden kann, sind periodische Indexe endlose Projekte,

[13] Fugmann 1999:218
[14] Cleveland/Cleveland 1990:27
[15] Cleveland/Cleveland 1990:33

die meistens von vielen Personen manchmal über mehrere Jahre hinweg durchgeführt werden. Konsistenzerhaltung wird dabei zu einer echten Herausforderung. Jede Ausgabe eines periodischen Index kann ganz unzusammenhängende Themen von verschiedenen Autoren in unterschiedlichen Stilen für sehr verschiedene Benutzer vereinen. Es gibt zwei Arten von periodischen Indexen: individuelle Indexe für individuelle Zeitschriften und breiter gefasste Indexe für eine Gruppe von Zeitschriften. Im ersten Fall entsteht unter der Leitung eines Herausgebers einer Zeitschrift ein Index, üblicherweise für das Zeitschriftvolumen eines Jahres. Diese Indexe sind wichtig und nützlich, obschon die breiter gefassten Indexe des zweiten Falls, die eine ganze Gruppe von Zeitschriften erfassen und normalerweise von Indexierungsdienstleistern gemacht werden, eine grössere Rolle spielen.

Die im nachfolgenden Abschnitt vorgestellten Indextypen sind alle innerhalb der soeben aufgeführten Indexierungsebenen denkbar.

2.3 Indextypen

Der primäre Faktor zur Unterscheidung von verschiedenen Indextypen ist die Anordnung der Einträge. Fast alle Standardindexe sind entweder alphabetisch oder klassifiziert angeordnet oder treten als Kombination der beiden Anordnungsvarianten auf. Die Disposition eines Index beeinflusst direkt die Benutzer des Index, denn die Suchenden können mit einem Index nicht umgehen, wenn sie dessen Struktur nicht verstehen. Eine schlechte Anordnung der Einträge eines Index beeinträchtigt die Effizienz und die Effektivität einer Suche. Folgende Indextypen können unterschieden werden:

Autorenindexe (*author indexes*)
Indexe, deren Einträge aus Namen von Personen, Koautoren, Organisationen, Regierungsabteilungen, Universitäten usw. bestehen, nennt man *Autorenindexe*. Ein Benutzer wird von einem solchen Index mittels Autorennamen zu den Titeln von Dokumenten geleitet. Im folgenden Beispiel wird auf die von Autoren verfassten Artikel in Fachzeitschriften verwiesen:

Aalto, A.	Wat. Pwr. Dam Constr., 33 (Mai 81) S. 40-3
Aaras, A.	Ergonomics, 23 (Aug 80) S. 707-26
Aarsten, M.	Design (Jun 81) S. 30-3
Aartsen, M.	Design (Mai 81) S. 36-7
Aastrup, S.	J. Inst. Brew., 86 (Nov-Dez 80) S. 277-83
Aatre, V. K.	IEE Proc. F: Commun. Radar Signal Process., 128 (Apr 81) S. 74-82
Abate, A.	Proc. Instn. Civ. Engrs. 2, 71 (Jun 81) S. 395-406
Abbady, M.A.	J. Chem. Technol. Biotechnol., 31 (Feb 81) S. 111-14

Autorenindexe sind sehr verbreitet und auch sehr nützlich. Sogar als indirekte Suche nach einem bestimmten Thema können Autorenindexe verwendet werden. Denn die in einer Branche arbeiteten Leute sind normalerweise vertraut mit den führenden Autoren ihres Gebietes und können sich so der Literatur mithilfe dieser bekannten Autorennamen nähern. Für die Erstellung von Autorenindexen ist es wichtig, dass genaue Richtlinien zur Verfügung stehen, damit die Konsistenz der Einträge und des Layouts gewährleistet ist.[16]

Die Autorenindexe sind nicht zu verwechseln mit einem Index, der von einem Textautor konstruiert wird, z.B. ein Schlagwortregister für das selbst verfasste Buch.

Alphabetische Schlagwortindexe oder **Schlagwortregister** (*alphabetical subject indexes*)
Der Ausdruck *alphabetischer Index* deckt eine ganze Anzahl von verschiedenen Arten von Indexen ab. In einem alphabetischen Index sind alle Themenausdrücke, Autorennamen, Ortsbestimmungen usw. alphabetisch angeordnet. Berücksichtigt werden müssen auch Symbole und Zahlen, die üblicherweise vor den Buchstaben am Anfang eines Index eingeordnet werden.[17] Die alphabetische Sortierung eines Index ist die üblichste Methode, da sie sehr bequem ist und sich jeder Benutzer damit auskennt. Aber es ist nicht die einzige Art von Anordnung. Ein Index kann eine klassifizierte Anordnung aufweisen oder gleichzeitig sowohl alphabetisch wie auch klassifiziert sein. Generell verlangt jedoch jede klassifizierte Anordnung eine alphabetische Zugangsmöglichkeit, um die Benutzung des Index effizient zu gestalten. Das Beispiel zeigt einen kurzen Auszug aus einem alphabetischen Schlagwortindex in einem Buch:

Hominiden 63
Homonymisierung 195
Hyperonyme/Hyperonymie 136, 172, 176
Input 25, 46, 91, 267-283, 295, 320-325
    defizitär 18
    negativ vs. positiv 21
Inputanpassungen 279, 288
Imitation 303, 312

Alphabetische Indexe haben für ihre Benutzer viele Vorteile. Sie begegnen uns jeden Tag und sind uns sehr vertraut, wenn wir beispielsweise ein Telefonbuch benutzen oder etwas in einem Wörterbuch nachschlagen. Es braucht für die Benutzung keine ausführliche Gebrauchsanweisung. Ein alphabetisch angeordneter Index kann durch einfaches Überfliegen zu einem Resultat führen und ist sehr zeitsparend. Probleme tauchen bei diesem Indextyp auf, wenn es um synonyme oder verwandte Einträge geht, welche im ganzen Index weit voneinander verstreut sein können. Wenn man nach dem Begriff „Hauskatze“ suchen will, stellt sich die Frage, ob man nun unter „Haus“ oder

[16] Cleveland/Cleveland 1990:40
[17] Cleveland/Cleveland 1990:45

unter „Katze“ oder eventuell sogar unter irgendeinem anderen Ausdruck sucht. Um diese Schwierigkeiten zu überwinden, werden Querverweise in den Index eingefügt, z.B. „siehe“ oder „siehe auch“.[18]

Das alphabetische Anordnen ist kein deterministischer Prozess. Ausdrücke können Buchstabe bei Buchstabe oder Wort für Wort alphabetisiert werden. Leerschläge, Sonderzeichen, Zahlen, Abkürzungen etc. müssen sinnvoll behandelt werden. Trotzdem hat sich die Technik bereits über Jahrhunderte hin bewährt, und sorgfältig ausgearbeitete Regeln wurden entwickelt, um den Prozess zu standardisieren.[19] Wir werden in Abschnitt 3.5.4 darauf zurückkommen.

Klassifizierte Indexe (*classified indexes*)
Dieser Typ von Index ist nicht nur in spezialisierten Gebieten, sondern auch für ganz allgemeine Themen beliebt. Ein klassifizierter Index ist für generelle Suchen sehr geeignet, wenn das Ziel der Suche ganze Klassen von Dokumenten sind. Ein klassifizierter Index ist als Hierarchie von miteinander verknüpften Themen angeordnet, beginnend mit allgemeinen Gegenständen und sich hinunterarbeitend zu den spezifischen. Es ist offensichtlich, dass ein Schema der Hierarchie bereits vor der Indexierung vorhanden sein muss.

Ein Ausschnitt aus einem klassifizierten Index im Bereich der Medizin könnte folgendermassen aussehen:

> A FIBERS see NERVE FIBERS
>
> ABATTOIRS
> Prevalence of disease conditions and pregnancy in sheep and goats seen over a three year period in Enugu abattoir, Nigeria. Wosu LO. Arch Roum Pathol Exp Microbiol 1988 Jan-Mar;47(1):57-64
>
> ABDOMEN
> Periarteritis with intra-abdominal bleeding. Sabo D et al. Ann Emerg Med 1988 Dec;17(12):1368
> Circadian rhythms in patients with abdominal pain syndromes. Roberts-Thomsen IC et al. Aust N Z J Med 1988 Jun;18(4):569-74
> Diagnostic investigation and special treatment of chronic abdominal ischaemia. Petrovsky BV et al. Int Angiol 1988 Jul-Sep; 7(3): 214-8

Klassifizierte Indexe haben Vor- und Nachteile. Der Hauptvorteil besteht darin, dass ein solcher Index auf konzeptuelle Weise eine Hilfe bei der Suche ist. Die Disposition basiert auf der Annahme, dass die Benutzer sich mit der logischen Anordnung von Wissen auskennen, und somit ist eine Suche dieser Art sehr natürlich. Wenn ein Benutzer einen Eintrag lokalisiert, so werden ihm auch automatisch andere mit dem Eintrag verknüpfte Einträge präsentiert. Man könnte den Prozess

[18] Obwohl es nach Ansicht vieler Leute keine echten Synonyme gibt, so haben Wörter doch oft fast gleiche Bedeutungen oder können ohne Kontext nicht eindeutig verstanden werden. Deshalb sollte Synonymie in einem Index mittels Querverweisen angezeigt werden. Cleveland/Ceveland 1990:77

[19] Cleveland/Cleveland 1990:47

vergleichen mit der Suche in einer nach Sachgebieten angeordneten Bibliothek, wenn man auch für die Bücher in der unmittelbaren Umgebung eines gesuchten Buches Interesse bekommt. Ein Nachteil ist sicherlich, dass zusätzlich zum klassifizierten Index auch ein alphabetischer Index zur Verfügung stehen muss, sodass eine Suche grundsätzlich ein zweistufiger Prozess ist, beginnend mit der Suche in der alphabetischen Liste, um dann die richtige Position in der klassifizierten Liste zu identifizieren. Ob ein klassifizierter Index oder eine Kombination sinnvoll ist, hängt ganz von den Bedürfnissen der Benutzer ab.

Koordinierte Indexe (*coordinate indexes*)
Koordinierte Indexe werden erstellt, indem zwei oder mehrere einzelne Terme durch Kombination derer einen neuen Eintrag kreieren. Werden z.B. die individuellen Ausdrücke „schwarz", „Hunde" und „Schweiz" kombiniert, so resultiert eine separate Klasse „schwarze Hunde in der Schweiz". Das Resultat ist eine Kombination der Einzelausdrücke. Dieser Indextyp stellt das Bool'sche Modell für eine Suche dar.

Wird die Koordination vom Benutzer zum Zeitpunkt der Suche vollzogen, so nennt man die Technik **Postkoordination** (*postcoordinate indexing*). Der Index wird auch als **manipulativer Index** (*manipulative index*) bezeichnet. Die Alternative ist, die einzelnen Indexterme bereits zum Indexierungszeitpunkt zu komplexen Themenbeschreibungen zusammenzusetzen und im Index zu verankern. In diesem Fall spricht man von **Präkoordination** (*precoordinate indexing*) oder von **nichtmanipulativen Indexen** (*nonmanipulative indexes*).[20] Grundsätzlich hat jeder Indexterm, ob post- oder präkoordiniert entstanden, das gleiche Gewicht. Keiner ist wichtiger als der andere.[21]

Der präkoordinierte Indexierungstyp ist für die traditionellen gedruckten Indexe von Bedeutung, da eine gedruckte Seite keine nachträgliche Manipulation erlaubt. Beispiele sind Buchindexe. Ein postkoordinierter Index behandelt jeden Term unabhängig von den anderen Termen und repräsentiert jeden Term einzeln als Eintrag im Index. Um die Bedürfnisse eines Benutzers zu erfüllen, werden bei der Recherche mehrere Terme mit Bool'schen Operatoren kombiniert. Das Problem dabei sind falsche Koordinationen, wenn ein Benutzer also etwas komplett anderes als das Gesuchte erhält, z.B. „Polish cars" bei einer Suche nach „car polish".

Permutierte Titelindexe (*permuted title indexes*)
In der Zeit, als Computer erstmals für die Informationsverarbeitung eingesetzt wurden, versuchten einige Leute, die Wörter der Dokumenttitel als Inhaltsindikator zu verwenden. Ein permutierter Titelindex geht von der Voraussetzung aus, dass ein Titel effektiv auf den Inhalt eines Dokuments verweisen kann. Und weil der Titel den Dokumentinhalt reflektiert, unterstützt ein permutierter Titelindex die Suchenden bei der Entscheidung, ob ein Dokument den Informationsbedarf

[20] Knorz 1997:124
[21] Cleveland/Cleveland 1990:61; Lancaster 1998:44

befriedigt. Permutierte Titelindexe werden kreiert, indem die informationstragenden Wörter in einem Titel systematisch rotiert bzw. permutiert und als Einträge in den Index aufgenommen werden. Selbstverständlich hängt diese Art von Index davon ab, wie gut ein Autor das Thema des Geschriebenen in dem Titel reflektiert. Der Titel wird dabei zur einzigen Komponente des Dokuments, wo ein Autor anzuzeigen versucht, worum es in dem Text geht.

Das beste Argument, das für die Schaffung von permutierten Titelindexen spricht, ist einfach: Die Erstellung ist einfach, mit einem Minimum an Kosten, vollständig automatisiert von einem Computer ausführbar. Die Nachteile sind: Die Titel widerspiegeln den Dokumentinhalt nicht immer angemessen. Die limitierte Anzahl der Wörter in den Titeln beschränken eine vollständige Gegenstandsanzeige. Die meisten dieser Indexe sind optisch nicht ansprechend und schwierig zu überfliegen. Der Mangel an Vokabularkontrolle kann die Menge der gefundenen irrelevanten Dokumente erhöhen. Eigentlich sind permutierte Titelindexe Konkordanzen und leiden somit auch unter den Schwächen derer, wie z.B. die Verstreuung von Synonymen und allgemeinen Begriffen. Zusätzlich gibt es bei permutierten Titelindexen keine oder nur eine geringe Konsistenzkontrolle, da die Indexe den Inhalt nur so weit reflektieren, wie die Wörter im Titel konzeptuelle Gegenstände repräsentieren. Die Effektivität von permutierten Titelindexen variiert je nach Themenbereich sehr stark.

Verschiedene Variationen von permutierten Titelindexen wurden entwickelt, um die Wirksamkeit zu verbessern. Beispielsweise versuchten Indexierer, den Titel mit zusätzlichen Wörtern zu ergänzen, welche sie vom Text selbst oder von einer Vokabularliste abgeleitet hatten oder von welchen die Indexierer glaubten, sie hätten mit den Wörtern im Titel etwas zu tun. Es wurden auch Querverweise zu verwandten Termen, besonders zu Synomymen, eingefügt. Jede dieser Techniken steigert aber die Kosten und den Zeitaufwand der Indexkonstruktion.

Die bekanntesten Formen von permutierten Titelindexen sind die sogenannten **KWIC-** und **KWOC-Indexe**.[22] Der erste Ausdruck bedeutet *key word in context* und der zweite *key word out of context*. Beide Bezeichnungen beziehen sich auf die Form der Indexeinträge.

Gegeben seien die folgenden Dokumenttitel:

Blauäugige Katzen in Texas
Die Katze und die Geige
Hunde und Katzen und ihre Krankheiten
Die Katze und die Wirtschaft

Der KWIC-Index ist eine alphabetische Liste, geordnet nach jedem inhaltstragenden Wort im Titel. Zur Identifizierung der inhaltstragenden Wörter werden häufig Stoppwortlisten verwendet (siehe

[22] Cleveland/Cleveland 1990:64; Lancaster 1998:48

Abschnitt 2.5.2.1.2). Für die obenstehenden Titel könnte ein KWIC-Index etwa diese Einträge haben (die Schlüsselwörter sind kursiv gedruckt, die Zahlen leiten den Benutzer zu den bibliografischen Daten der Dokumente):

in Texas,	*Blauäugige* Katzen	23
Die	*Katze* und die Wirtschaft	12
Die	*Katze* und die Geige	17
Hunde und	*Katzen* und ihre Krankheiten	3
Blauäugige	*Katzen* in Texas	23
und ihre	Krankheiten, Hunde und Katzen	3
ihre Krankheiten,	*Hunde* und Katzen und	3
und die	*Wirtschaft*, Die Katze	12
und die	*Geige*, Die Katze	17
in	*Texas*, Blauäugige Katzen	23

Im Gegensatz dazu rotiert ein KWOC-Index die Wörter nicht, er hebt jedoch die Schlüsselwörter hervor und listet sie separat auf:

Blauäugige	Blauäugige Katzen in Texas	23
Katze	Die Katze und die Wirtschaft	12
Katze	Die Katze und die Geige	17
Katzen	Blauäugige Katzen in Texas	23
Katzen	Hunde und Katzen und ihre Krankheiten	23
Krankheiten	Hunde und Katzen und ihre Krankheiten	3
Hunde	Hunde und Katzen und ihre Krankheiten	3
Wirtschaft	Die Katze und die Wirtschaft	12
Geige	Die Katze und die Geige	17
Texas	Blauäugige Katzen in Texas	23

Facettierte Indexe (*faceted indexes*)

Ein facettiertes Schema ist eine Art von synthetischer Klassifikation und wird oftmals **analytisch-synthetisches System** (*analytico-synthetic system*) genannt. Das System gestattet Postkoordination und ist in einer Klassifikationsordnung angeordnet, also nicht alphabetisch. Wird ein solches Schema zur Indexierung von Dokumenten verwendet, resultiert daraus ein facettierter Index. Jeder Term im Klassifikationssystem ist einer sogenannten **Facette** (*facet*) zugeordnet und steht in unterschiedlicher Relation zu den anderen Termen. Diese sind nicht vollständig formuliert (z.B. „air-to-ground missile“), sondern als Bausteine intendiert („air“, „ground“, „missile“), sodass auch das Konzept „ground-to-air-missile“ zulässig ist. Eine Facette kann auch als Baum verstanden werden, und eine facettierte Klassifikation als Sammlung von Bäumen. Jedem Knoten eines Baumes entspricht eine Klassifikationsmarke bzw. ein Indexterm. Das Gesamt der Baumstrukturen wird durch Verknüpfungen der Terme erweitert. Dokumente, die mithilfe einer solchen Klassifikation indexiert werden, können damit mehrere Facetteneltern besitzen. Sucht ein Benutzer beispielsweise Informationen über ein Portrait, so wird er durch den facettierten Index auch auf

Dokumente verwiesen, welchen die Indexterme „Vita“, „Lebensdaten“, „Biografie“, „Lebenslauf“ usw. zugeordnet wurden.

Facettenanalyse ist ein streng kontrollierter Prozess, bei welchem einfache Konzepte in sorgfältig definierten Kategorien organisiert sind, indem Klassennummern der Basiskonzepte verbunden werden. Sie stehen im Gegensatz zu den **aufzählenden** oder **enumerativen** Klassifikationssystemen[23], die meistens vor der Anwendung mithilfe von Präkombinationen vorstrukturiert werden. Aufzählende Systeme sind starr und können neues Wissen nur auf begrenzte Art unterbringen, während facettierte Systeme flexibler sind.[24]

Facette bezeichnet nach Fugmann eine Gruppe von Begriffen, welche einer begrifflichen Kategorie des Klassifikationssystems oder eines Thesaurus zugeordnet sind.[25] Eine Kategorie ist der oberste Allgemeinbegriff, für den es auf dem betreffenden Fachgebiet keinen sinnvollen, noch allgemeineren Oberbegriff gibt. Facette bedeutet auch eine Seite von etwas, das viele Seiten hat. In Bezug auf Indexerstellung heisst das, dass jedes Thema nicht ein einzelnes Element ist, sondern viele Aspekte hat. Somit versucht ein Facettenindex, alle individuellen Aspekte eines Themas zu entdecken und diese künstlich herzustellen auf eine Art, die das Thema am besten beschreibt.

Wie jeder andere klassifizierte Index benötigt auch ein Facettenindex einen alphabetischen Index, um seinen Gebrauch zu erleichtern. Je grösser und komplexer ein Index wird, desto nötiger wird ein zusätzlicher alphabetischer Index für die schnelle Lokalisierung bestimmter Konzepte im Index.

Kettenindexe (*chain indexes*)

Indexbenutzer riskieren, nützliche Einträge zu verpassen, wenn sie bei ihrer Suche nicht nach dem spezifischsten Thema ihres Interesses suchen. Kettenindexierung ist eine Methode, die dieses Risiko zu minimieren versucht, indem die einzelnen Einträge eines klassifizierten Index einer nach dem andern in einer alphabetischen Liste präsentiert werden. Das heisst, Kettenindexe ermöglichen die Verknüpfung oder die Verkettung jedes Konzepts mit den direkt verwandten Konzepten in dem Hierarchiesystem. Die Terme werden von allgemein zu spezifisch verkettet, und alle Terme oder Gegenstandsbezeichnungen, für die ein Element indexiert werden kann, werden mit einbezogen.[26] Wird ein Dokument beispielsweise mit den Termen „Elefanten“, „Menschen“ und „dick“ indexiert, so ist nicht klar, ob das Dokument von Elefanten und dicken Menschen oder von dicken Elefanten und Menschen handelt. Ein Kettenindex versucht, die Beziehungen zwischen den einzelnen Indextermen aufzuzeigen, und würde etwa folgenden Eintrag beinhalten:

[23] Fugmann 1999:76
[24] Cleveland/Cleveland 1990:65
[25] Fugmann 1999:215
[26] Lancaster 1998:54

Elefanten
dick Menschen

Daraus kann ein Benutzer ersehen, dass es im Dokument um dicke Elefanten und ihre Beziehung zu Menschen geht. Dicke Menschen in Bezug auf Elefanten würden im Index an einer anderen Stelle auftreten.

Kettenindexierung überwindet das Problem der allgemeinen **Eingangs-** oder **Zugangseinträge** in einem klassifizierten Index, indem sie alle signifikanten Terme als Eingangseinträge benutzt. Solche *entry terms* sind Stellen, die einem Benutzer den Zugang zur Indextermliste erlauben.[27] Die Indexierung ist sehr mechanisch und entlastet den Indexierer von einer Menge von Entscheidungen. Auf der anderen Seite handelt es sich beim Kettenindex um ein klassifiziertes System, welches sowohl die Fähigkeiten wie auch die Mängel eines klassifizierten Systems reflektiert.

Zeichenkettenindexe (*string indexes)*
Eine Zeichenkette (*string*) ist üblicherweise, jedoch nicht zwangsläufig, eine Ausgabe eines Computers. Die Idee eines Zeichenkettenindex ist es, eine Reihe von rotierten Indexeinträgen aus einer Basisliste von Indextermen, welche die Zeichenketten darstellen, zu entfalten. Das Ziel dabei ist, einem Benutzer einen Eingangseintrag für alle Indexterme zur Verfügung zu stellen und diese mit ihrem Kontext zueinander darzustellen.

Nehmen wir an, es geht in einem Text um das Thema „Der Gebrauch von Computeranimationen für das Fernsehen in den Vereinigten Staaten“. Eine Textanalyse ergibt einen linearen String von Indextermen „Computer. Animation. Fernsehindustrie. Vereinigte Staaten.“. Diese Terme werden dann von einem Computer unter Berücksichtigung der Beziehungen zwischen den Termen rotiert, und es wird eine Reihe von Einträgen produziert:

VEREINIGTE STAATEN
Fernsehindustrie. Computer. Animation.

FERNSEHINDUSTRIE. Vereinigte Staaten.
Computer. Animation.

COMPUTER. Fernsehindustrie. Vereinigte Staaten.
Animation.

ANIMATION. Computer.
Fernsehindustrie. Vereinigte Staaten.

[27] Cleveland/Cleveland 1990:87

Zeichenkettenindexierung ist eine Form von computerunterstütztem Indexieren. Intellektuelle Entscheidungen werden vom Indexierer getroffen, die zeitraubende Manipulationsarbeit wird von einer Maschine ausgeführt. Gegenwärtig werden Forschungen unternommen, um herauszufinden, bis zu welchem Grad ein Computer immer mehr der manuellen Aktivitäten übernehmen kann.[28] Das phänomenale Wachstum von online Datenbankinformationssystemen und die begleitende Entwicklung des computerunterstützten Indexierens begünstigen ein kontinuierliches Interesse in die Zeichenkettenindexierung. Es stehen den Indexierern eine ganze Anzahl von Softwareprodukten zur Verfügung und neue erscheinen regelmässig.

Zitierindexe (*citation indexes*)
Ein Zitierindex (zuweilen auch Zitatindex genannt) besteht aus einer Liste von Dokumenten mit einer Unterliste zu jedem Dokument von nachfolgenden publizierten Arbeiten, die das Dokument zitieren:

> Joshi, Aravind K. et al. „Tree Adjunct Grammars." Journal of Computer and System Sciences, vol. 10, no. 1 (1975), pp. 136-163.
>
> > Harbusch, Karin. „The Relation between Tree-Adjoining Grammars and Constraint Dependency Grammars." MOL5 (1997), pp. 38-45.
> >
> > Kroch, Anthony S. et al. „Analyzing Extraposition in a Tree Adjoining Grammar." Syntax and Semantics, vol. 20 (1987), pp. 107-49.

Die Idee von Zitierindexen ist nicht neu; bereits vor mehr als einem Jahrhundert entwickelte der Berufszweig der Juristen einen solchen Index, den „Shepard's Citations".[29] Es ist offensichtlich, dass gerade für diese Berufsgruppe ein solches Werkzeug fundamental ist. Einen Zitierindex jedoch als generelles Referenzwerkzeug zu benutzen, ist ein eher neuer Ansatz.

Zitierindexierung ist ein komplett andersartiger Ansatz des Indexierens. Er hängt in keiner Weise von den Indexwörtern ab und kennt somit keine Probleme im Zusammenhang mit Bedeutungen und Interpretationen von Bedeutungen, die beim konventionellen Indexieren spürbar sind. Beim Erstellen von Zitierindexen sind die Autoren selbst die qualifizierten Personen, die das für ihr Sachgebiet relevante Material definieren und schildern. Weil ein Zitierindex nicht auf der Zuweisung von Indextermen basiert und keine Interpretation oder Terminologie benötigt, besteht zwischen dem Autor und dem Benutzer ein klarer Verständigungskanal, auch ohne die Einführung einer künstlichen Sprache. Ein Zitierindex impliziert, dass ein zitiertes Dokument eine innere

[28] Cleveland/Cleveland 1990:69
[29] Cleveland/Cleveland 1990:72

Themenbeziehung zum Dokument hat, welches es zitiert, und dieser Bezug wird für das Gruppieren von Dokumenten benutzt.

Der primäre Vorteil beim Gebrauch eines Zitierindex ist derjenige, dass er den Benutzer zu den neusten Artikeln führt; das heisst, dass ein Zitierindex in der Zeit vielmehr vorwärts als rückwärts geht. Bis zu einem gewissen Punkt kann ein Benutzer auch ausfindig machen, wie Ideen eines bestimmten Dokuments entstanden sind und welche Entwicklungen sich später daraus ergeben haben. Ein anderer Vorteil liegt darin, dass nur der Autor selbst in die Beurteilung von „was ist relevant“ involviert ist. Und somit kann die Produktion vollständig automatisiert werden. Die zwei offensichtlichen Nachteile von Zitierindexen sind die hohen Kosten für die Produktion eines solchen Index sowie die Abhängigkeit davon, dass die Grundlage eines solchen Index auf der Annahme basiert, dass die Autoren der Dokumente konsistent sind und gute Kenntnisse über ihre Zitate haben.

Zitierindexe sind mehr als nur Werkzeuge zur Informationssuche. Sie bilden auch eine Forschungsgrundlage für das Studium der Verhaltenscharakteristiken der Literaturkonsumenten sowie -produzenten und für die Struktur des Wachstums der Wissenschaft selbst.[30]

Damit haben wir die wichtigsten Typen von Indexen kennen gelernt. Wir gehen nun zur Sprache der Einträge von Indexen über.

2.4 Die Indexsprache

Für jeden Index ist die Sprache der Indexeinträge von äusserster Wichtigkeit. Mittels sprachlicher Ausdrücke findet ein Benutzer an einer entsprechenden Stelle die gesuchte Information in einem Text wieder. Dazu muss nicht nur der Indexierer seine Einträge mit sprachlichen Elementen formulieren, auch der Benutzer muss sein Informationsbedürfnis in einer sprachlich formulierten Frage ausdrücken. Sind die gewählten Ausdrücke oder deren Zusammenhänge mangelhaft, misslingt eine Suche. Aufgrund dieser Tatsachen widmet sich der folgende Abschnitt der Indexsprache, der künstlichen Sprache also, welche mit ausreichender Treffsicherheit die Essenz von eingespeicherten Texten aufzufinden gestattet und das Ziel des zuverlässigen Wiederfindens von gespeicherter Information verfolgt. Die Begriffe *Indexsprache*, *Indexierungssprache*, *Dokumentationssprache* und *Retrievalsprache* werden in der Literatur üblicherweise synonym verwendet.

[30] Cleveland/Cleveland 1990:73

2.4.1 Der Indexterm

Ein **Indexterm** (*index term*) ist ein sprachlich formulierter Zugangspunkt oder ein Identifikator eines Textes, über welchen der Text oder eine Stelle im Text lokalisiert und wieder auffindbar gemacht werden kann.[31] Da diese Definition nicht für alle Indexterme der verschiedenen Indextypen exakt passt, sollten entsprechend die folgenden Typen von Indextermen unterschieden werden:

Schlüsselwörter (*keywords*) sind Wörter, die direkt aus dem zu indexierenden Text stammen. Bei Fugmann und Kaiser werden diese auch ***Stichwörter*** genannt.[32]

Deskriptoren (*descriptors*) sind Terme, die aus einem Thesaurus stammen.

Notationen (*notations*) sind Indexterme, die aus einem Klassifikationssystem stammen.

Identifikatoren (*identifiers*) sind in einem Index verwendete Eigennamen, z.B. Personennamen, Organisationen, Projektnamen, Ortsnamen, Firmennamen usw. Identifikatoren beschreiben einzigartige Entitäten und nicht generelle Konzepte. Wenn sie in Texten erscheinen, müssen sie in derselben Ausdrucksweise (evtl. in normalisierter Form) auch im Index auftreten.

Schlagwörter (*subject index terms*) bezeichnen zum einen die Indexeinträge, die in Buchregistern vorkommen. Zum andern gelten als Schlagwörter diejenigen Terme, die aus einer Schlagwortliste stammen (siehe Abschnitt 2.4.2.2.3). Schliesslich existiert auch eine weiter gefasste Vorstellung von Schlagwort, die ganz allgemein alle Terme, die irgendeinem Dokument als Indexterme zur Inhaltserschliessung zugeordnet werden, als Schlagwörter definiert. Zu dieser Art von Schlagwörtern gehören dann auch die Deskriptoren und Notationen. Charakteristisch für alle Schlagwörter ist, dass sie im Dokument selbst, welchem sie zugewiesen werden, nicht unbedingt vorhanden sein müssen und dass ein Schlagwort aus einem Einzelwort oder auch einer lexikalischen Wortgruppe bestehen kann. Ich lege mich hier auf diejenige Definition fest, die ein Schlagwort als einen in einem Buchregister vorkommenden Indexterm identifiziert, unabhängig davon, woher das Schlagwort kommt. Zur Bezeichnung der Terme, die irgendeinem Dokument während des Indexierungsprozesses zugeordnet werden, benutze ich vorzugsweise den Begriff *Indexterme*.

Alle diese Indextermtypen sind Teil des Indexsprachenwortschatzes, des Vokabulars also, aus welchem die Einträge in einem Index gebildet werden. Wir werden uns im folgenden Abschnitt mit dem Indexsprachenwortschatz und danach mit der Indexsprachengrammatik beschäftigen.

[31] Moens 2000:50
[32] Fugmann 1999:221; Kaiser 1996:4

2.4.2 Der Indexsprachenwortschatz

Um das Wesen eines Indexterms näher zu bestimmen, werden wir zunächst einige begriffliche Unterscheidungen vornehmen.

2.4.2.1 Begriffsdefinitionen

Innerhalb des Vokabulars, das zur Erstellung eines Index verwendet wird, gilt es, die beiden Begriffe *Allgemeinbegriff* und *Individualbegriff* zu unterscheiden.[33] Ein **Allgemeinbegriff** ist ein Begriff, welchem man noch weitere sinnvolle Merkmale hinzufügen kann. Unter **Begriff** selbst versteht man die Summe aller wesentlichen Aussagen, die man über einen Gegenstand machen kann (ich werde für *Begriff* auch synonym den Ausdruck *Konzept* verwenden). Jede Aussage bildet ein **Merkmal** des Begriffs. Ein Allgemeinbegriff beschreibt also immer eine Klasse von miteinander verwandten Gegenständen, von Gegenständen nämlich, die eine Menge von wesentlichen Merkmalen miteinander gemeinsam haben. Fügt man dem Allgemeinbegriff ein weiteres Merkmal hinzu, dann führt das zu einem spezifischeren Unterbegriff.

Im Gegensatz zum Allgemeinbegriff ist es für einen **Individualbegriff** typisch, dass bei ihm sämtliche begrifflichen Merkmale eines Gegenstandes festgelegt sind und dass ihm kein weiteres Merkmal mehr hinzugefügt werden kann, zumindest keines, welches aus der Perspektive eines betreffenden Fachgebietes noch sinnvoll wäre. Beim Individualbegriff ist die Spezifizierung durch Hinzufügen von sinnvollen Merkmalen an eine Grenze gestossen.

Auf dieser Unterscheidungsgrundlage erweisen sich die Individualbegriffe als einfach handhabbar im Indexierungsprozess, ganz im Gegensatz zu den Allgemeinbegriffen. Die Suche nach Individualbegriffen, z.B. die Suche nach einem ganz bestimmten Buch, wird *known item search* genannt.

Für das Verständnis der Gesetzmässigkeiten, die in einem Informationssystem herrschen, bedarf es neben der Unterscheidung von Individual- und Allgemeinbegriffen auch einer Differenzierung von der lexikalischen und der nichtlexikalischen Ausdrucksweise. In Anlehnung an Fugmann definieren wir als **lexikalisch** eine jede fest vereinbarte, lineare Zeichenfolge, in welcher ein Begriff bzw. eine Begriffsverknüpfung ausgedrückt ist.[34] Es kann sich hierbei um ein Einzelwort handeln oder auch um eine aus mehreren Wörtern bestehende Ausdrucksweise. Der Wortlaut muss allerdings festgelegt sein (wie z.B. bei „Rotes Kreuz“). Die lexikalischen Ausdrucksweisen können in einem

[33] Fugmann 1999:18-26
[34] Fugmann 1999:29

Lexikon an einem festen und voraussehbaren Platz eingeordnet und an einem solchen Platz gezielt wieder aufgefunden werden. Bei gewissen lexikalischen Ausdrücken ist nicht sicher, dass jedermann weiss, was gemeint ist, zumal sie in der Fach- und Umgangssprache auch oftmals in unterschiedlichen Bedeutungen gebraucht werden. Der korrekte Umgang mit lexikalischen Ausdrücken erfordert deshalb häufiges Nachschlagen in entsprechenden Wörterbüchern zur Bedeutungsklärung.

Zu den **nichtlexikalischen** Ausdrucksweisen werden alle frei formulierten, definitionsartigen, paraphrasierenden und erklärenden Ausdrücke für einen Begriff in Gestalt von Sätzen und Nominalphrasen gerechnet.[35] Diese paraphrasierende Ausdrucksweise steht für einen jeden Begriff sofort zur Verfügung, und zwar auch ohne dass hierfür erst besondere sprachliche Vereinbarungen und Konventionen getroffen werden müssen. Man kann eine solche Ausdrucksweise jederzeit verwenden, um deutlich und unmissverständlich zu sagen, was man meint, und braucht hierzu nicht erst die Prägung einer lexikalischen Ausdrucksweise abzuwarten. Die nichtlexikalische Ausdrucksweise ist sogar bisweilen die einzige Ausdrucksmöglichkeit für neue Begriffe überhaupt, zumindest in der Anfangszeit. Selbst dann, wenn bereits eine lexikalische Ausdrucksweise existiert, wird oftmals die nichtlexikalische bevorzugt. Denn die nichtlexikalisch-paraphrasierende Ausdrucksweise erfordert keinerlei Nachschlagen zur Vergewisserung ihrer Bedeutung.

Es gibt sehr viele Varianten, wie man auf nichtlexikalische Weise einen Begriff zum Ausdruck bringen kann. So ist die Mannigfaltigkeit der Ausdrucksweisen, in denen uns die Allgemeinbegriffe in der natürlichen Sprache begegnen oder begegnen können, praktisch unbegrenzt gross. Das Auftreten von Allgemeinbegriffen in der paraphrasierenden Ausdrucksweise ist von grösster Bedeutung für die Gestaltung und Handhabung eines Informationssystems. Denn in der Bevorzugung der nichtlexkalischen Ausdrucksweise für Allgemeinbegriffe liegt nach Auffassung von Fugmann ein gravierendes Hindernis für die Nutzung der Wörter in einem Text für dessen Indexierung.[36] Individualbegriffe hingegen werden fast immer in der lexikalischen Ausdrucksweise angetroffen. Hierdurch ist die Vielfältigkeit, in welcher Individualbegriffe in der natürlichen Sprache auftreten, begrenzt. Durch diesen Umstand wird die Suche nach Individualbegriffen in einem Informationssystem drastisch erleichtert.

Auch im Zusammenhang mit Indexsprachwortschätzen gilt es, nochmals die **Präkombination** (*precombination*) von der **Postkombination** (*postcombination*) zu unterscheiden (vergleiche Abschnitt 2.3).[37] Präkombination oder Präkoordination herrscht, wenn ein Indexterm die Bedeutung von mindestens einem anderen Indexterm mit umfasst, welcher bereits im Wortschatz vorhanden ist oder die Aufnahme in dieses Vokabular beanspruchen könnte. Bereits zum Indexierungszeitpunkt werden einzelne Indexterme zu komplexen Themenbeschreibungen kombiniert und im Index verankert. Beispiele für Präkombination sind die Begriffsbenennungen in der Bedeutung von

[35] Fugmann 1999:30
[36] Fugmann 1999:33
[37] Fugmann 1999:53; Knorz 1997:124

Rosten (als Verknüpfung von „Eisen“ und „Korrosion“) oder Luftkühlung (wenn die Begriffe „Luft“, „Kühlen“, „Eisen“, „Korrosion“ ebenfalls im Wortschatz vertreten sind oder Aufnahme beanspruchen können). Von Postkombination oder Postkoordination spricht man, wenn die Kombination von Indextermen zum Zeitpunkt einer Recherche zugelassen ist. In einem gedruckten Buchregister ist Postkombination grundsätzlich nicht möglich. Jeglicher Begriff, welcher mehr als einen einzigen Begriff kategorialen Charakters umfasst, wie im Beispiel, erweist sich als präkombinierend oder **polykategorial**. Ein Begriff, in welchem dies nicht der Fall ist, gilt als **Einfachbegriff** oder **monokategorialer Begriff**.[38]

2.4.2.2 Kontrolliertes Vokabular

Beim Wortschatz oder Vokabular einer Indexierungssprache ist ein kontrolliertes von einem freien Vokabular zu unterscheiden.[39] Ein **kontrolliertes Vokabular** (*controlled vocabulary*) ist eigentlich eine Autoritätsliste. Das heisst, dass Indexierer einem Dokument nur diejenigen Terme zuweisen dürfen, die in dieser Liste stehen. Diese beispielsweise von einer Indexagentur oder einem Herausgeber erstellte Liste sollte konsistent sein. Üblicherweise ist das kontrollierte Vokabular mehr als einfach nur eine Liste, sondern ein Netz mit irgendeiner Art von semantischer Struktur. Solche Netze haben verschiedene Aufgaben. Erstens einmal soll die generelle Konzeptstruktur in einem Fachgebiet repräsentiert werden. Sie gewährleisten auch die Kontrolle von Synonymen, indem eine Standardform ausgewählt und ein Bezug zu den synonymen Formen hergestellt wird. Auch Terme, deren Bedeutungen eng verwandt sind, werden in einem Netz verknüpft. Dabei gibt es zwei Arten von Beziehungen: eine **hierarchische** und eine nichthierarchische oder **assoziative**. Mittels Querverweisen werden horizontale und vertikale Verhältnisse aufgezeigt. Ferner definiert ein solches Vokabular – falls nötig – mehrdeutige Ausdrücke, und es unterscheidet Homographe. Denn die meisten Wörter der natürlichen und der fachspezifischen Sprache sind mehrdeutig.

Eine weitere wichtige Aufgabe eines kontrollierten Vokabulars ist die Vergrösserung der konzeptuellen Übereinstimmung zwischen dem Indexierer und dem Benutzer. Ein Indexierer untersucht einen Text, filtert mental die Absichten des Autors und wählt dann die Indexterme aus einem kontrollierten Vokabular, welche die angemessenen, geeigneten Konzepte und Bezüge anhand der Interpretation des Indexierer darstellen. Der Benutzer nähert sich dem Index mit persönlichen Konzepten und Begriffen. Die Funktion des kontrollierten Vokabulars ist es, sowohl den Indexierer als auch den Benutzer am Schluss zu demselben Punkt zu führen. Vokabularkontrolle ist laut Cleveland/Cleveland ein komplexes Gebiet und erfordert eine der wichtigsten und grundlegendsten Entscheidungen beim Indexieren, nämlich die Entscheidung für die Verwendung eines kontrollierten oder eines freien Indexwortschatzes.[40]

[38] Fugmann 1999:53
[39] Cleveland/Cleveland 1990:77; Knorz 1997:125; Lancaster 1998:19
[40] Cleveland/Cleveland 1990:77

Die drei Haupttypen von kontrollierten Vokabularen und Hilfsmitteln beim Indexieren sind **Thesauri** (*thesauri*), **Klassifikationssysteme** (*classification schemes*) und **Schlagwortlisten** (*list of subject headings, subject authority files*).[41] Fugmann ergänzt die drei Typen ausserdem um den **Classaurus** (*classaurus*), einer Mischung aus Thesaurus und Klassifikation.[42] Alle vier Typen versuchen, die potenziellen Indexterme alphabetisch und systematisch zu präsentieren. Sie werden nun genauer erläutert.[43]

2.4.2.2.1 Thesaurus

Der Thesaurus als Variante eines Indexsprachenwortschatzes wird bei Burkhart in seinen wesentlichen Merkmalen beschrieben und ist im informationswissenschaftlichen Sinne analog zu DIN 1463 so definiert:

> „Ein Thesaurus im Bereich der Information und Dokumentation ist eine geordnete Zusammenstellung von Begriffen und ihren (vorwiegend natürlichsprachigen) Bezeichnungen, die in einem Dokumentationsgebiet zum Indexieren, Speichern und Wiederauffinden dient. Er ist durch folgende Merkmale gekennzeichnet:“[44]

a) Begriffe und Bezeichnungen werden eindeutig aufeinander bezogen (terminologische Kontrolle), indem Synonyme möglichst vollständig erfasst werden; Homonyme und Polyseme werden besonders gekennzeichnet, und für jeden Begriff wird eine Bezeichnung (Vorzugsbenennung, Begriffsnummer oder Notation) festgelegt, die den Begriff eindeutig vertritt.

b) Beziehungen zwischen Begriffen (repräsentiert durch ihre Bezeichnungen) werden dargestellt.

c) Der Thesaurus ist präskriptiv, indem er für seinen Geltungsbereich festlegt, welche begrifflichen Einheiten zur Verfügung gestellt werden und durch welche Bezeichnungen diese repräsentiert werden.

Etwas allgemeiner formuliert ist ein Thesaurus ein für die Indexierung und Recherche vorgegebenes Vokabular von natürlichsprachlichen Begriffsbenennungen, welches zugleich auch hierarchische und assoziative Verwandtschaften zwischen den Begriffen darstellt.[45]

---

[41] Cleveland/Cleveland 1990:77; Knorz 1997:125; Lancaster 1998:15
[42] Fugmann 1999:78
[43] Es gibt Fälle, da stehen auch (Online-) Terminologiedatenbanken als Hilfsmittel für den Prozess des Indexierens zur Verfügung. Sie werden hier nicht weiter besprochen. Für weitere Informationen dazu sei auf die Arbeit von Niederbäumer 2000 verwiesen.
[44] Burkart 1997:160
[45] Fugmann 1999:64

Die Hauptfunktion eines Thesaurus ist die Verallgemeinerung oder Uniformierung von Termen, die verwandte Bedeutungen haben, jedoch unterschiedliche Oberflächenformen. Er hat im Detail die folgenden Aufgaben:

1. Eine erste wichtige Aufgabe ist die Synonymkontrolle. Synonyme Wörter werden durch Wortsubstitution behandelt. Ein Thesaurus bringt synonyme und austauschbare Wörter innerhalb einer **Äquivalenzklasse** zusammen. Wenn die natürliche Sprache mehrere Ausdrücke für das gleiche oder fast das gleiche Konzept hat, so führt der Thesaurus das Vokabularangebot üblicherweise zu einem einzigen rechtsgültigen Term.

2. Falls ein Thesaurus hierarchische Beziehungen zwischen den Wörtern enthält, kann er für eine Erweiterung der Terme verwendet werden. In diesem Fall wird ein Wort eines Originaltextes durch den breiter gefassten Klassenterm ersetzt. Für die Suche nach allgemeinen Konzepten ist das sehr nützlich. Gelegentlich kann ein Thesaurus die Bedeutung von Termen auch einengen.

3. Viele Wörter der natürlichen Sprache haben mehr als nur eine semantische Bedeutung. Ein Thesaurus kann diese Mehrdeutigkeiten auflösen. Die Technik beinhaltet die Verwendung des Wissens um die syntaktische Klasse eines Wortes, das indexiert wird, sowie des Fachwissens, das eine Wortklasse mit einer Wortbedeutung verbindet.[46]

Was den Aufbau eines Thesaurus betriff, um also von der natürlichen Sprache als Ausgangsmaterial zum kontrollierten Vokabular eines Thesaurus zu gelangen, müssen mehrere kontrollierende und definierende Prozesse durchlaufen werden. Zuerst muss der Bezugsrahmen eingegrenzt werden, denn ein Thesaurus kann den Anforderungen bezüglich Eindeutigkeit, Verbindlichkeit und Übersichtlichkeit nur dann gerecht werden, wenn der Kontext (*universe of discourse*), den er abdecken soll, klar umrissen ist. Alle bisherigen Versuche eines universalen Thesaurus müssen „als fehlgeschlagen oder nicht vollendet betrachtet werden“.[47]

Die folgenden Elemente des Bezugsrahmens sollten zu Beginn der Thesaurusarbeit abgesteckt werden:

- Gegenstandsbereich oder Thematik des Thesaurus (Schwerpunkte, Randgebiete)
- Spezifität des Thesaurus (bis zu welcher Spezifität oder bis zu welchem Allgemeinheitsgrad sollen Begriffe einbezogen werden)
- Sprachstil des Thesaurus (mehr wissenschaftlich orientiert oder auch für Nichtfachleute verständlich)
- Umfang des Thesaurus (Umfang des Vokabulars, Umfang der ausgewiesenen Begriffsbeziehungen und Beziehungsarten)

[46] Moens 2000:107
[47] Burkart 1997:163

Für diese Parameter können keine allgemeingültigen Angaben gemacht werden, vielmehr hängen sie vom Benutzerkreis, den zu erschliessenden Dokumenten und der angestrebten Erschliessungstiefe ab.

Erst wenn dieser Rahmen festgelegt ist, kann mit der eigentlichen Erarbeitung des Thesaurus begonnen werden, mit der Wortgutsammlung. Dazu müssen die Quellen ausgewählt werden, denen Wortgut entnommen werden kann. In der erstellten Wortgutsammlung werden dann durch **terminologische Kontrolle** die Mehrdeutigkeiten und Unschärfen aufgelöst. Hierzu sind gemäss Burkart drei Schritte nötig: Synonymkontrolle, Polysem- und Homonymkontrolle sowie Zerlegungskontrolle (Kompositazerlegung).

Durch die **Synonymkontrolle** werden mehrere Bezeichnungen zu einer begrifflichen Einheit zusammengefasst. Beim Indexieren werden Synonymbeziehungen mit „siehe“-Verweisen ausgedrückt. Mit der **Polysem-** und **Homonymkontrolle** werden Bezeichnungen, die mehrere begriffliche Einheiten beinhalten, entsprechend dieser Einheiten auf Bezeichnungen aufgeteilt, die voneinander differieren. Durch die **Zerlegungskontrolle** wird versucht, einen für den Gegenstandsbereich des Thesaurus angemessenen Spezifitätsgrad der begrifflichen Einheiten zu erreichen und zusätzliche sprachliche Einstiegsmöglichkeiten in den Thesaurus zu schaffen. Der grosse Nachteil einer Zerlegung ist offensichtlich: Der Thesaurus wird sehr umfangreich und unübersichtlich. Als Gegenposition dazu gibt es das Uniterm-Verfahren. Dabei versucht man, alle Komposita zu vermeiden und Wörter zu verwenden, die nicht weiter in Bedeutungsbestandteile zerlegbar sind. Diese Einheiten oder elementaren Basisbegriffe nennt man **Uniterms**.[48]

Die durch die terminologische Kontrolle entstandenen begrifflichen Einheiten werden **Äquivalenzklassen** (*equivalence classes*) genannt. Für die Behandlung und Darstellung der Äquivalenzklassen gibt es verschiedene Möglichkeiten. In einem **Thesaurus ohne Vorzugsbenennung** werden alle Elemente der Äquivalenzklasse gleich behandelt und können unterschiedslos für Indexierung und Retrieval genutzt werden. Die Äquivalenzklasse wird in diesem Fall von einer Begriffsnummer repräsentiert, die das Bindeglied zwischen den verschiedenen Bezeichnungen bildet. In einem **Thesaurus mit Vorzugsbenennung** wird ein Klassenelement jeder Äquivalenzklasse als Vorzugsbenennung ausgewählt. Dieses ausgewählte Element nennt man **Deskriptor** (*descriptor*). Alle anderen Elemente haben den Status von Nichtdeskriptoren oder Synonymen. Sie werden in den Thesaurus aufgenommen, bilden einen Bestandteil des Zugangsvokabulars, können aber selbst nicht zur Indexierung und Recherche verwendet werden, sondern verweisen auf den entsprechenden Deskriptor. Ein Deskriptor sollte seine Äquivalenzklasse möglichst umfassend, zweifelsfrei und präzise darstellen, am Sprachgebrauch des Fachgebietes orientiert sein sowie einprägsam und möglichst unkompliziert sein. Je besser diese Kriterien erfüllt werden, desto selbsterklärender ist der Thesaurus und kann auf zusätzliche Erläuterungen wie *scope*

[48] Burkart 1997:167; Knorz 1997:124

notes verzichten. In einer *scope note* oder **Erläuterungskategorie** werden Hinweise zum spezifischen Gebrauch eines Deskriptors festgehalten.

Neben der terminologischen Kontrolle werden mittels **begrifflicher Kontrolle** Bedeutungsverschiebungen zwischen Bezeichnungen der natürlichen Sprache und den Deskriptoren eines Thesaurus dargelegt. Bei dieser Kontrolle treten Beziehungen zwischen Begriffen auf, die vielfältiger Natur sein können. In Thesauri beschränkt man sich meistens auf drei Beziehungsarten oder Relationen: Äquivalenzrelation, hierarchische Relation, Assoziationsrelation.[49]

Die **Äquivalenzrelation** ist eine innerbegriffliche Relation zwischen Bezeichnungen. Sie verweist in einem Thesaurus mit Vorzugsbenennung von einem Nichtdeskriptor zum dazugehörigen Deskriptor und von einem Deskriptor zu allen entsprechenden Nichtdeskriptoren. Durch die **hierarchische Relation** wird die begriffliche Ober- und Unterordnung ausgedrückt. Dabei werden zwei Typen von hierarchischer Relation unterschieden: die generische Relation, die eine Relation definiert zwischen zwei Begriffen, von denen der Unterbegriff alle Merkmale des Oberbegriffs besitzt, und die partitive Relation, welche eine Relation zwischen zwei Begriffen, von denen der übergeordnete Begriff (Verbandsbegriff) einem Ganzen entspricht und der untergeordnete Begriff (Teilbegriff) einen der Bestandteile des Ganzen repräsentiert. Die **Assoziationsrelation** ist eine zwischen zwei Bezeichnungen als wichtig erscheinende Beziehung, die weder eindeutig hierarchischer Natur ist noch als äquivalent angesehen werden kann. Die Beziehungen, die hier verankert sind, haben ganz unterschiedlichen Charakter und stehen in einem assoziativen Sinne zueinander. So gehört beispielsweise „zum Rosten“ intuitiv auch „der Rost“ und „das Rostschutzmittel“. Häufig werden in diese Beziehungsart alle Relationen hineingepackt, die bei den anderen zwei Arten nicht richtig hineinpassen.[50] In einem Index sind Assoziationsrelationen durch „siehe auch“-Verweise ausgedrückt.

Ist die Möglichkeit der Begriffskombination gegeben, so ergibt sich ein weiterer Relationstyp: Der **zusammengesetzte Begriff**, der im Thesaurus durch die Kombination von zwei Deskriptoren wiedergegeben werden soll, ist formal eigentlich ein Nichtdeskriptor. Er gehört jedoch nicht den Äquivalenzklassen der Deskriptoren an, durch die er gebildet wird, vielmehr wäre seine Äquivalenzklasse eine Schnittmenge der beiden Äquivalenzklassen der Deskriptoren. Diese existiert im Thesaurus aber nicht a priori, sondern wird erst bei einer Kombination gebildet.

Für die sinnvolle Darstellung eines Thesaurus sind ebenso Überlegungen anzustellen; diese gewährleistet nämlich eine optimale Verwendung. Auch wenn die meisten Thesauri auf einem Rechner geführt werden, wird parallel dazu oft eine gedruckte Ausgabe bereitgehalten. Für beide Arten bieten sich alphabetische und/oder systematische Anordnung an.

[49] Burkart 1997:171; Fugmann 1999:66-96
[50] Fugmann 1999:26f.

Einer der wichtigsten Punkte beim Erstellen und Verwenden eines Thesaurus ist dessen Unterhalt und Pflege. Im Gegensatz zur natürlichen Sprache muss in einem Thesaurus jeder Sprachwandel angepasst werden. Regelmässige Kontrollen und Revisionen sind unerlässlich. Wird der Unterhalt eines Thesaurus vernachlässigt, so sind alle Vorzüge eines kontrollierten Vokabulars schnell verschwunden.[51]

Aufgrund der Ausführungen in diesem Abschnitt ist klar geworden, dass sowohl der Aufbau wie auch die Pflege eines Thesaurus ausserordentlich aufwändig sind. Um Kosten und Zeit zu sparen, wird das Bedürfnis nach maschinell erstellten Thesauri immer grösser. Allerdings ist das keine leichte Aufgabe. Besonders schwierig scheint die automatische Definition einer Beziehungsart zwischen Termen zu sein.[52]

2.4.2.2.2 Klassifikation

Unter Klassifikation wird ganz allgemein eine Gruppierung oder Einteilung des gesamten Wissens der Wissenschaft und ihrer Disziplinen nach einheitlichen methodischen Prinzipien verstanden.[53] Die Elemente (Bestandteile) der Klassifikation bezeichnet man als **Klassen**.[54]

Manecke empfiehlt, bei der Verwendung des Begriffes *Klassifikation* zu unterscheiden zwischen:

- dem Prozess der Klassifikationserarbeitung (d.h. der Klassenbildung),
- dem Klassifikationssystem als Ergebnis des Klassenbildungsprozesses,
- dem Prozess der Klassierung bzw. des Klassifizierens, d.h. dem gegenseitigen Zuordnen von Objekten und Klassen des Klassifikationssystems.[55]

Ein Klassifikationssystem ist insgesamt das Ergebnis eines schrittweisen Strukturierungsprozesses, bei dem jeder Klasse in dem System ein bestimmter Platz zugeteilt wird. So erfüllen die Klassifikationen vor allem eine Ordnungsfunktion (Gleiches zu Gleichem). Die begriffliche Über- und Unterordnung ist in einer Klassifikation deutlich erkennbar.

Hinter jeder Klasse verbirgt sich ein dreistufiger Abstraktionsprozess, also zunächst die Abstraktion vom Objekt bzw. Sachverhalt einer Klasse zum Begriff, der die Merkmale bestimmt (vergleiche Abschnitt 2.4.2.1), die diese Masse von einer anderen unterscheidet. Dieser Begriff ist dann in einer nächsten Abstraktionsstufe durch eine äquivalente **Bezeichnung** auszudrücken.

---

[51] Cleveland/Cleveland 1990:94; Moens 2000:108f.
[52] Moens 2000:132
[53] Manecke 1997:141
[54] Lancaster 1998:17
[55] Manecke 1997:141

Innerhalb von Klassifikationssystemen werden üblicherweise zwei Bezeichnungsarten nebeneinander verwendet. Zum einen sind das die verbalen in der Systematik zusammengeführten Begriffe (**Benennungen** genannt). Diese Bezeichnungen können auch aus mehreren Wörtern zusammengesetzt sein. Im Interesse einer praktikablen und widerspruchsfreien Handhabung von Klassifikationen fordert Manecke sorgfältig gewählte und verständliche Bezeichnungen.[56] Ergänzt werden die Bezeichnungen – falls nötig – durch Erläuterungen oder Definitionen. Die andere Bezeichnungsart sind künstliche Bezeichnungen in Form von **Notationen** (*notations).* Sie entsprechen inhaltlich den Begriffsbezeichnungen und bilden insgesamt ein die jeweilige Klassifikation repräsentierendes und charakterisierendes Notationssystem. Bei der Inhaltserschliessung eines Textes werden die Notationen als inhaltskennzeichnende Merkmale (**Indizes**) vergeben. Sie dienen auch als Grundlage für das Retrieval. Die Notationen bilden bei einer Indexierung mittels Klassifikation den Wortschatz der Indexsprache.

Das bekannteste Beispiel einer Klassifikation ist die Internationale Dezimalklassifikation. Sie ermöglicht einen sachgebietsorientierten Zugang zu allen Wissensgebieten und ist eine auf der Zehnerzahlenteilung basierende, im Wesentlichen monohierarchische Universalklassifikation. Die Notationen werden mit Dezimalzahlen dargestellt, welche zwecks einer besseren Lesbarkeit durch Punkte gegliedert sind.[57] Die Colon-Klassifikation ist etwas weniger verbreitet. Sie ist eine teilfacettierte Universalklassifikation mit alphabetisch repräsentierten Notationen.[58] Das nachfolgend dargestellte Beispiel ist der Internationalen Dezimalklassifikation entnommen:

als natürlichsprachlicher Ausdruck:	als Notation:
„Einfahrt und Ausfahrt bei Garagen und Parkplätzen“	656.052.468

In einer Klassifikation liegt eine starke Hierarchie (**Monohierarchie**) vor, wenn zu jedem Begriff mehrere Unterbegriffe existieren. Es entsteht eine Art Begriffspyramide, bei der jeder Artbegriff umgekehrt nur einen Oberbegriff hat. Eine Suche innerhalb der Klassifikation ist hier nur nach einem Aspekt möglich und somit eindimensional. Eine schwache Hierarchie (**Polyhierarchie**) liegt vor, wenn ein und derselbe Begriff aufgrund der Berücksichtigung mehrerer unterschiedlicher Merkmale jeweils zwei oder mehr Oberbegriffen zugeordnet wird. Das ermöglicht eine gleichzeitige Recherche unter mehreren Aspekten oder eine mehrdimensionale Suche.

Bei der Entwicklung und Anwendung von Klassifikationen unterscheidet Manecke zwischen zwei Typen, den **analytischen** und den **analytisch-synthetischen Systemen**[59]. In einer typischen, starr

[56] Manecke 1997:144
[57] Die Internationale Dezimalklassifikation geht auf Melvil Dewey (1851-1931) zurück. Er veröffentlichete 1876 die erste Ausgabe seiner Dewey Decimal Classification (DDC). Seither wurde diese auf internationaler Ebene systematisch weiterentwickelt und in diverse Sprachen übersetzt. Ferber 2000:1
[58] Manecke 1997:148-156
[59] Manecke 1997:145

strukturierten analytischen Klassifikation werden die in der Systematik zusammengeführten Begriffe entsprechend den Gegebenheiten des Fachgebietes von oben nach unten, vom Allgemeinen zum Speziellen, immer feiner untergliedert. Bei der Vergabe von Indizes darf nur das verwendet werden, was in der Klassifikation präkoordiniert enthalten ist. Im Gegensatz dazu gehen analytisch-synthetische Klassifikationen (auch als *Facettenklassifikationen* bezeichnet) von den in einer Systematik zusammengestellten, gleichrangigen Merkmalsbegriffen eines Sachgebietes aus (z.B. Objekte, Eigenschaften, Personen, Zeit), welchen entsprechende Einzelbegriffe zugeordnet werden. Diese Einzelbegriffe werden **Foci** oder **Isolate** genannt. Die auf diese Art entstandenen Begriffsgruppen bezeichnet man als **Kategorien** oder **Facetten**. Die notwendige Untergliederung erfolgt durch weitere (Unter-)Facetten. Analytisch-synthetische Klassifikationen sind in der Regel ahierarchisch und mehrdimensional. Mit ihnen können postkoordinierend auch sehr komplexe Sachverhalte durch eine Synthese von Begriffen aus verschiedenen Facetten wiedergegeben werden.

Klassifikationen gelten im Gegensatz zu Thesauri als relativ ausdrucksschwache Indexsprachen, da von den Möglichkeiten der Begriffsbeziehungen (Äquivalenz, Hierarchie, Assoziation) meist nur die hierarchischen Beziehungen das Systemgefüge bestimmen. Zum Teil werden Assoziationen durch Querverweise ermöglicht.[60] Den Klassifikationen wird auch vorgeworfen, dass sie zu breit gefasst sind und zu weite Territorien abdecken, um den spezifischen Bedürfnissen der Indexsprachen gerecht werden zu können. Ferner mangelt es ihnen an präzisen Definitionen.[61] Die dennoch grosse Verbreitung von Klassifikationen in der Praxis der Indexierungsarbeit beruht vor allem auf drei vorteilhaften Eigenschaften:

- Universalität: die Orientierung auf den gesamten Bereich der Wissenschaft (als Universalklassifikation) bzw. auf viele ihrer Teilgebiete (als Fachklassifikation).
- Kontinuität: die Verwendbarkeit über einen längeren Zeitraum hinweg und ihre Fähigkeit zur Berücksichtigung neuer Erkenntnisse zur Erhaltung von Aktualität.
- Flexibilität: die Erweiterungsmöglichkeiten (auch als **Hospitalität** bezeichnet) durch Möglichkeiten zur Streichung von Begriffen und zur Bildung neuer Begriffsklassen sowie die **Expansivität**, also die Fähigkeit, das Klassifikationssystem in unterschiedlichen Gliederungsebenen darzustellen und zu benutzen.[62]

Die in Deutschland gültigen Regeln zur Erarbeitung und Weiterentwicklung von Klassifikationssystemen sind festgelegt in der Norm DIN 32705: Klassifikationssysteme. Erstellung und Weiterentwicklung von Klassifikationssystemen.

[60] Manecke 1997:144
[61] Cleveland/Cleveland 1990:85
[62] Manecke 1997:146

2.4.2.2.3 Schlagwortliste

Eine traditionelle Schlagwortliste ist eine Liste mit Schlagwörtern, die durch Querverweise verbunden sein können. Sie ist analog zu einem Thesaurus alphabetisch geordnet und unterscheidet sich von einem Thesaurus wegen des Fehlens einer vollständigen hierarchischen Struktur und wegen des Mangels an einer klaren Differenzierung von hierarchischen und assoziativen Verhältnissen. Dies sind auch die Gründe dafür, dass Schlagwortlisten für den Indexierungsprozess kaum verwendet werden.[63]

2.4.2.2.4 Classaurus

Eine sinnvolle Vokabulargestaltung besteht gemäss Fugmann in einer Kombination der Konzepte von Thesaurus und Klassifikation.[64] Man legt dem Wortschatz die systematisch-hierarchische Struktur einer Klassifikation zugrunde und arbeitet mit schlagwortartigen Begriffsbenennungen, sofern diese in der natürlichen Sprache existieren. Dadurch werden erstens der Mangel an Mnemonik bei den Notationen vermieden und zweitens der Mangel an Aufnahmefähigkeit behoben, welcher jedem Thesaurus anhaftet. Drittens wird bestmögliche Aktualität gesichert.

Diese Idee einer Zusammenführung von Thesaurus und Klassifikation ist in einer Variante erstmals von Bhattacharyya unter der Bezeichnung **Classaurus** beschrieben worden. Auf diese Weise verfügt ein Indexierer über die spezifischen Vorteile von Thesaurus und Klassifikation, und zugleich lassen sich ihre Mängel weit gehend vermeiden.[65]

Nach diesen Ausführungen zu den gängigsten Werkzeugen beim Indexieren mit einem kontrollierten Vokabular werden nun einige Vor- und Nachteile der Verwendung eines solchen Vokabulars aufgezeigt.

Ein kontrolliertes Vokabular hat auf der einen Seite viele Vorteile und vermeidet viele Probleme offener Systeme. Die Form eines Indexterms sowie dessen Bedeutung sind festgelegt, Mehrdeutigkeiten werden dadurch aufgelöst. Durch die Festlegung entsprechen die Terme denjenigen Begriffen, die in einer Fach- oder in der Umgangssprache benutzt werden. Ein kontrolliertes Vokabular ermöglicht es einem Indexierer, eine Benennung zu finden, an die er sich vorübergehend nicht erinnern kann. Beziehungen wie Synonymie, Hierarchie und Assoziation können auf einfache Weise in den Index übernommen werden. Insgesamt wird die Indexierung konsistenter und effizienter.

---

[63] Lancaster 1998:15
[64] Fugmann 1999:78
[65] Fugmann 1999:79

Auf der anderen Seite gibt es für jedes Mehr an Kontrolle in einem Vokabular auch Probleme und Nachteile. Ein kontrolliertes Vokabular ist naturgemäss auf diejenigen Begriffe beschränkt, für welche lexikalische Ausdrücke existieren. Ein Indexierer muss sich an das Vokabular halten, auch wenn ihm einmal ein natürlichsprachliches Wort, das nicht im Indexvokabular auftritt, adäquater erscheint. Das Indexieren mit einem kontrollierten Vokabular benötigt einen grösseren Zeitaufwand und verursacht höhere Kosten sowohl beim Erstellen des kontrollierten Indexvokabulars als auch beim Prozess des Indexierens. Die Analyse von Indexierungsfehlern zeigt, dass bei Verwendung umfangreicher Thesauri oder Klassifikationen beispielsweise auch der erfahrene und spezialisierte Indexierer ohne Nachschlagen in keinem Falle auskommt. Hinzu kommt, dass für den Unterhalt und die Pflege grosse Anstrengungen nötig sind, die sich in hohen Kosten und grossem Zeitaufwand niederschlagen. Ein kontrolliertes Indexvokabular leidet fast immer an Aktualitäts- und Spezifitätsmangel. Ausserdem ist der Erarbeitungs- und Erprobungsaufwand enorm. Der Zwang zu Bedeutungsklärung und die Suche nach bestpassenden Begriffsbenennungen sind nicht immer nur von Vorteil. Und ein Fehlen von Begriffsbenennungen darf unter keinen Umständen auftreten, sofern dies überhaupt möglich ist.

2.4.2.3 Freies Vokabular

Unter Indexieren mit einem **freien** oder **unkontrollierten Vokabular** (*free or uncontrolled vocabulary*) versteht man die Verwendung frei gewählter natürlichsprachlicher Ausdrücke als Indexterme. Die Formulierung der freien Indexterme sollte sich an die vorgefundene Formulierung im Text anlehnen sowie prägnant, reproduzierbar und eindeutig sein. In DIN 31623 („Indexierung zur inhaltlichen Erschliessung von Dokumenten“) sind detaillierte Richtlinien für die Begriffs- und Benennungsanalyse bei der Formulierung freier Indexterme angegeben.[66]

Bei der Wahl von freien Indextermen sind einige Hürden zu bewältigen: Ein Wort kann aufgrund der paraphrasierenden Ausdrucksweise viele Begriffe bedeuten und ein Begriff durch viele grundverschiedene Wörter oder Wortgruppen ausgedrückt werden. Beim Indexieren muss eine Begriffsbenennung so formuliert werden, dass alle Mehrdeutigkeiten aufgelöst werden. Besonders schwierig wird es bei Begriffen, für die es (noch) keine lexikalische Bezeichnung gibt, oder bei Allgemeinbegriffen. Ganz allgemein lässt sich festhalten, dass die unbegrenzte Mannigfaltigkeit der natürlichen Sprache einem Indexierer und selbstverständlich auch dem Suchenden grosse Schwierigkeiten bereiten kann.

Als Vorteil von freien Indexvokabularen kann die Freiheit bei der Wortwahl gesehen werden. Es kann jedoch auch vorkommen, dass die von einem Indexierer definierten Indexterme sich im

[66] Knorz 1997:125

Verlaufe des Indexierungsprozesses verändern und spezifischer oder allgemeiner werden. Im schlimmsten Fall entwickelt sich sogar ein komplettes Chaos, sodass der Index unbrauchbar wird.[67]

Die Nachteile eines freien Vokabulars sind offensichtlich: Inkonsistenz, Mangel an Spezifität, Mangel an Relationsbildungen, geringe Voraussehbarkeit der Ausdrucksweise, Missverständlichkeit, Lückenhaftigkeit usw. Ein grosser Anteil der Last wird bei dieser Art von Indexierung auf den Benutzer gelegt, denn er muss durch Raten eruieren, wie er zum Ziel kommen könnte. Zusätzlich hängt das Suchziel beim Retrieval stark von der Wortwahl des Autors ab. Die der natürlichen Sprache innewohnende Unvorhersehbarkeit der von den Autoren gewählten Ausdrucksweisen für Begriffe wird möglicherweise zu einem grossen Hindernis für einen vollständig formulierten Suchauftrag. Gerade beim Suchen mit Allgemeinbegriffen ist es äusserst schwierig, ein vollständiges Recherchenergebnis zu erzielen.

Es herrscht aufgrund all dieser Punkte in der Sekundärliteratur Einigkeit darüber, dass ein kontrolliertes Vokabular bei einer sorgfältigen Indexierung von natürlichsprachlichen Texten unabkömmlich ist. Allerdings wird auch eine Kombination von kontrolliertem und freiem Vokabular empfohlen, um einigen Nachteilen der kontrollierten Vokabulare entgegenwirken zu können, z.B. wenn keine adäquaten Deskriptoren zur Verfügung stehen.[68]

2.4.2.4 Die Anforderungen an ein Indexierungsvokabular

Es ist nach Ansicht von Lancaster nicht in erster Linie die Art des Vokabulars, also frei oder kontrolliert, welches eine Indexerstellung beeinflusst, sondern die Spezifität eines Vokabulars.[69] In diesem den Gegenstand des Indexsprachenwortschatzes abschliessenden Abschnitt werden deshalb Anforderungen aufgezeigt, die ein Indexierungsvokabular erfüllen sollte.

- **Korrektheit**: Ein Indexvokabular soll keine Fehler enthalten.
- **Spezifität**: Bei jedem Indexsprachenwortschatz ist eine möglichst hohe Spezifität der Indexterme anzustreben, die auch zukünftigen Anforderungen gewachsen sein sollte. Mit Präkombination kann höhere Spezifität erreicht werden.
- **Aktualität**: Ein aktuelles und jederzeit erweiterbares Vokabular ist für den Erfolg der Indexerstellung unerlässlich.
- **Exhaustivität**: Ein Vokabular soll innerhalb eines Fachbereichs so vollständig wie möglich sein und keine Erfassungslücken aufweisen.
- **Lexikalisierung**: Der Indexsprachenwortschatz stellt Begriffsbenennungen für paraphrasierende Ausdrucksweisen zur Verfügung.

[67] Cleveland/Cleveland 1990:79
[68] Cleveland/Cleveland 1990:79; Fugmann 1999:99; Knorz 1997:125
[69] Lancaster 1998:19

- **Bedeutungsfestlegung**: Ein Vokabular legt die genauen Bedeutungen von Indextermen fest, um Mehrdeutigkeiten aufzulösen.
- **Begriffsanalyse**: Um auch nach den einzelnen Bestandteilen eines Indexterms recherchieren zu können, stellt das Indexsprachenvokabular die Bestandteile des Terms zur Verfügung (z.B. „Kupfer“ und „Zink“ bei „Messing“).
- **Terminologie- und Begriffskontrolle**: Synomymie, Polysemie, Homonymie, Hyponymie und Hyperonymie sowie assoziative Begriffsrelationen werden durch Querverweise erfasst.
- **Ellipsenauffüllung**: Ein natürlichsprachlicher Text weist in seiner Ausdrucksweise zur Vermeidung von Wiederholungen immer Lücken auf. Diese lexikalischen Lücken sollten im Index aufgefüllt sein.
- **Voraussehbarkeit und Rekonstruierbarkeit**: Die Indexterme eines Vokabulars müssen auch zu einem zukünftigen Zeitpunkt voraussehbar und rekonstruierbar sein.

2.4.3 Die Indexsprachengrammatik

Bei einer **gleichordnenden Indexierung** (*coordinate indexing*) fehlt jede Art von (syntagmatischen) Beziehungen zwischen Indextermen. Die Indexierung besteht aus einer Menge von gleichberechtigten Indexeinträgen. Mit einer derartigen Indexsprache, welche sich auf den Wortschatz beschränkt, lassen sich nach Ansicht von Fugmann stets nur bescheidene Recherchenergebnisse erzielen.[70]

Wann immer bei einer Indexierung zugeteilte Indexterme nicht völlig gleichberechtigt sind, sind syntaktische Mittel unverzichtbar, um die Relationen zwischen den Zuteilungen auszudrücken.[71] Aus diesem Grund hat sich die Bezeichnung **syntaktische Indexierung** (*syntactic indexing*) etabliert. Knorz zieht der Bezeichnung *syntaktische Indexierung* die seiner Auffassung nach treffendere Bezeichnung **strukturierte Indexierung** (*structured indexing*) vor.[72] Die beiden Bezeichnungen werden in dieser Arbeit synonym verwendet, denn obwohl der Ausdruck *strukturierte Indexierung* treffender sein mag, so hat sich *syntaktische Indexierung* im Fachbereich des Indexierens eingebürgert.

Als strukturierte Indexierung der einfachsten Form kann bereits ein Text verstanden werden, der im Freitext unter Verwendung von Kontextoperatoren recherchierbar ist (siehe Abschnitt 2.5.2.1.1). Jedes bedeutungstragende Wort ist dabei ein suchbares Stichwort, und so besteht die Indexierung aus einer Folge (Sequenz) von Indextermen. Die Nachbarschaftsbeziehungen der einzelnen Wörter bleiben erhalten und können bei der Suche berücksichtigt werden. Eine andere Interpretation einer Reihung liegt vor, wenn Indexterme nach Wichtigkeit geordnet sind. Vor allem automatische

[70] Fugmann 1999:79
[71] Knorz 1997:125
[72] Knorz 1997:125f.

Indexierungsverfahren ordnen den Deskriptoren oder Schlagwörtern oft **Gewichte** (*weights*) zu, die eine weitere Differenzierung darstellen (siehe Abschnitt 2.5.2.1.3.1). Mehrere Indexterme können durch syntaktische Mittel zu einer neuen Einheit zusammengesetzt werden, um einen gemeinten Begriff präziser zu benennen. Mit einer **Klammerstruktur**, welche die Abhängigkeitsstruktur zwischen Indextermen nachzeichnet, lässt sich das Gemeinte im Index klarer darstellen. Speziell bei Registerindexierungen bildet man zweckmässigerweise Paare von Indextermen, wobei der eine Term (Untereintrag, *qualifier*) den anderen (Schlagwort, *main heading*) präzisiert (vergleiche Abschnitte 3.2.1.1 und 3.2.1.2).

Die Wirkungsweise einer Indexsprachensyntax lässt sich bildlich wie in Abbildung 1 darstellen.

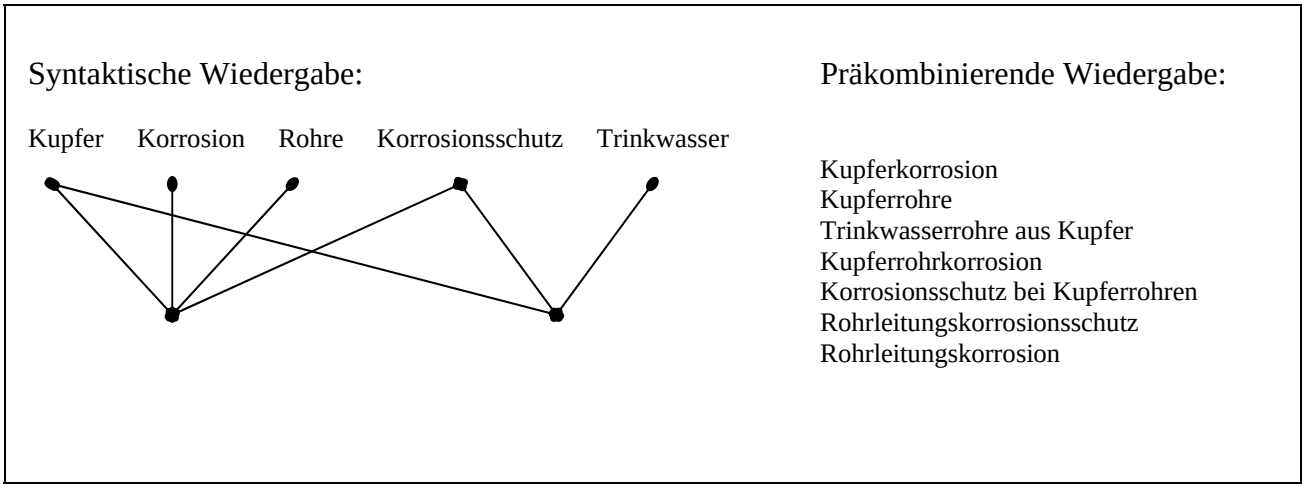

Abbildung 1: Syntaktische und präkombinierende Wiedergabe von Indextermen

Auf der linken Seite ist durch Verbindungsstriche angedeutet, wie man Begriffe syntaktisch miteinander verknüpfen könnte. In der rechten Darstellungshälfte finden sich eine Reihe von Präkombinationen, die man auf einem grösseren Spezialgebiet bei einer grammatiklosen Indexsprache wohl benötigen würde, um eine ausreichend exakte Indexierung und Recherchengenauigkeit zu erzielen. Alle Präkombinationen wären beim Einsatz einer wirkungsvollen Indexsprachengrammatik überflüssig und könnten dort ersatzlos aus dem Indexsprachenwortschatz gestrichen werden. Das Ergebnis ist ein viel kleinerer, viel langsamer wachsender und besser übersichtlich bleibender Wortschatz.[73] Denn ein Wortschatz, zu welchem Präkombinationen ungehindert Zugang haben, kann enorm schnell wachsen, und das Netzwerk der Begriffsverknüpfungen wird immer undurchschaubarer.

Der Aufwand bei der syntaktischen Indexierung besteht in einer fortgesetzten, kategoriengesteuerten Begriffsanalyse bei der Indexierung sowie auch in der Durchführung dieser

[73] Fugmann 1999:80

Begriffsanalyse bei der Recherche. Ein weiteres ernstes Hindernis für den Einsatz einer Indexsprachengrammatik dürfte die Abneigung des nichtprofessionellen Nutzers sein, sich mit den Regeln einer solchen künstlichen Grammatik und mit der Begriffsanalyse vertraut zu machen.[74]

Die am meisten verbreiteten genutzten syntaktischen Werkzeuge sind laut Fugmann die Segmentbildung (bei Datenbanken), die Schlagwortkette (im Bibliothekswesen), die Relationenindikatoren und die topologischen Verfahren.[75]

2.4.3.1 Die Segmentbildung

Damit beispielsweise der Zusammenhang zwischen den einzelnen Bestandteilen einer Legierung zur Suchbedingung gemacht werden kann, muss man bei der Indexierung mehr tun, als lediglich zusammenhanglos alle Legierungsbestandteile aufzuzählen, die in einem Dokument genannt sind. Vielmehr muss der Zusammenhang zwischen den Legierungsbestandteilen beschrieben werden, und zwar so, dass man ihn auch als Suchbedingung formulieren kann. Dieser enge Zusammenhang kann mit dem syntaktischen Hilfsmittel der **Segmentbildung** ausgedrückt werden.

Ein **Segment** ist eine vom Indexierer gebildete Gruppe von Begriffsbenennungen innerhalb eines Dokuments, zwischen denen ein besonders enger, im Dokument postulierter Zusammenhang der nichthierarchischen Art besteht.[76] Segmente lassen sich auch für vielerlei andere Arten von Begriffsverknüpfungen einrichten, z.B. für das parasitische oder symbiotische Zusammenleben von Lebewesen, für das Auftreten von Stoffen in Lebewesen oder in Organen derselben usw.

Wesentlich für die Segmentierung ist, dass klar definiert wird, welche Art von Begriffsverknüpfung in einem Segment ausgedrückt wird. Ein Segment muss dann die betreffende Kennzeichnung tragen, z.B. „LEG“ für ein Legierungssegment. Dann kann man die Recherche mithilfe dieser Kennzeichnung z.B. auf die Legierungssegmente beschränken und findet dann die Metalle in genau dieser Art von Verknüpfung vor. Ein Segment für die Verknüpfung von Stoff und Stoff-Attribut würde man vielleicht mit „SA“ kennzeichnen.[77]

Die Segmentbildung ist ein sehr wirksames und einfaches syntaktisches Werkzeug. Man kann auf diesem Weg die Zukunftssicherheit eines grossen und rasch weiterwachsenden Informationssystems gewährleisten. Bei der Indexierung verursacht die Segmentbildung aber zusätzlichen Aufwand. Sie muss erlernt und fortlaufend ausgeübt werden. Auch der Fragesteller muss über die entsprechenden Kenntnisse verfügen, wenn er dieses Hilfsmittel nutzen will. Für kleine Informationssysteme ist es fraglich, ob sich der Aufwand lohnt.

[74] Fugmann 1999:81
[75] Fugmann 1999:81-93
[76] Fugmann 1999:81
[77] Fugmann 1999:83

2.4.3.2 Die Schlagwortkette

Anstelle von Notationen lassen sich nach einem ebenfalls vorgegebenen Muster auch Schlagwörter und Deskriptoren zusammenfügen. Damit erzielt man eine ähnliche Verbesserung der Spezifität und Recherchengenauigkeit wie beim Zusammenfügen von Notationen, ebenfalls unter Inkaufnahme eines Mehraufwandes bei der Indexierung.

Eine **Schlagwortkette** wird dadurch erzeugt, dass bei der Indexierung die einzelnen Schlagwörter nicht lediglich in nichtinterpretierter Weise niedergeschrieben werden. Es wird vielmehr auch die Art ihres Zusammenhanges dargestellt. Hierzu ordnet der Indexierer jedem Schlagwort einen bestimmten Rollenoperator zu, je nach der besonderen Rolle, die ein Gegenstand bei einem Vorgang spielt. Dieser **Rollenoperator** oder **Rollenindikator** (*role indicator*) weist den Gegenstand z.B. als den aktiven oder passiven Teilnehmer an einem Vorgang aus, als den Teil eines Systems von Gegenständen oder als die Eigenschaft eines solchen, als geografische Lokalität usw.[78] Hier ein Beispiel von Rollenindikatoren:

Italien (A): Ausfuhr nach Italien
Italien (B): Einfuhr von Italien

Die syntaktische Verknüpfung der Schlagwörter, die auf diesem Wege erreicht wird, dient nicht nur dazu, einem Leser die Zusammenhänge vor Augen zu führen. Sie kann vielmehr auch als Suchbedingung bei der maschinellen Recherche genutzt werden. Für den Zugriff mittels Computer spielt die Reihenfolge der einzelnen Schlagwörter in einer Schlagwortkette nur eine untergeordnete Rolle, denn ein Schlagwort ist für das Programm an jeder beliebigen Stelle in seiner Kette auffindbar, nicht nur dann, wenn es als Ordnungswort an den Anfang der Schlagwortkette gerückt ist.

2.4.3.3 Der Relationenindikator

Ein weiteres syntaktisches Hilfsmittel ist das Prinzip der **Relationenindikatoren**, das auf der Grundlage der Segmentbildung basiert. Dabei gibt es für jeden Vorgangsbegriff eine ganze Familie von Begriffen, die besonders eng mit dem Vorgang zusammenhängen, jedoch auf nichthierarchische, assoziative Weise. So existieren beispielsweise für den Vorgang der Korrosion das korrosiv wirkende Mittel und auch das Material, welches der Korrosion unterliegt (aktive und passive Vorgangsteilnehmer). Es gibt die Vorgangsbeschleuniger und die Vorgangsverzögerer.

[78] Knorz 1997:127

Daneben steht das Ergebnis des betreffenden Vorganges, beim Eisen also beispielsweise der Rost, beim Kupfer der Grünspan.

Ein einfacher Weg zu einer Lexikalisierung mithilfe von Relationenindikatoren besteht darin, dass man an das Vorgangsschlagwort eine Ziffer anhängt, welche die Art des Familienmitglieds charakterisiert. Ein zwischengeschaltetes „+“ drückt beispielsweise die konfirmative Bedeutung aus, ein „–“ die negierende Bedeutung.

Korrosion+1	Korrosion (im Dokument ausdrücklich beschrieben)
Korrosion–1	Nichteintritt von Korrosion

Hat man ein vollständiges Vokabular für alle wesentlichen Vorgänge auf seinem Fachgebiet entwickelt, dann verfügt man auf diesem Weg zugleich auch über ein grösseres Vokabular an vorgangsbezogenen, assoziativ verwandten Indextermen. Hier ist man unabhängig davon, ob die Natur- oder Fachsprache für einen Begriff schon einen lexikalischen Ausdruck gebildet hat. Dies bedeutet, dass man dann praktisch für jede Art von Vorgangsbeteiligung eine Begriffsbenennung zur Verfügung hat. Die Genauigkeit der Indexierung lässt sich wesentlich steigern und damit zugleich auch die Genauigkeit der Recherche. Trotz der zahlreichen Vorteile der Relationenindikatoren sind sie dennoch bisher nur in einem Grossgebiet der Chemie und dessen Grenzgebieten erfolgreich angewandt worden.[79]

Die Relationenindikatoren gelangen zur vollen Wirkung, wenn man sie zusammen mit der Segmentbildung einsetzt, um die syntaktische Verknüpfung eines Gegenstandes mit der Art seiner Vorgangsbeteiligung zum Ausdruck zu bringen. Dann verfügt man über sehr präzise wirkende Suchmöglichkeiten.

2.4.3.4 Links und Proximity-Operatoren

Ein weiteres syntaktisches Werkzeug zur Erzeugung einer Indexsprachengrammatik stammt aus der Frühzeit der mechanisierten Dokumentation: das **Linking**. Hierbei wird alles, was von einem Autor in einem natürlichsprachlichen Satz (oder auch in einer Überschrift) zusammengefasst worden ist, als irgendwie zusammengehörig angesehen und mit einer gemeinsamen Zahl, dem **Link**, gekennzeichnet, wie im folgenden Beispiel aus Fugmann dargestellt:[80]

Leinöl[1] verharzt[1] nach Zusatz von Mangansalzen[1] im Tageslicht[1] und bei Zutritt von Luft[1] besonders schnell.

[79] Fugmann 1997:90
[80] Fugmann 1999:90

In der Fragestellung vertraut man dann darauf, dass ein gesuchter Zusammenhang zwischen mehreren Begriffen auch wirklich vom Autor eines einschlägigen Dokuments innerhalb eines einzigen Satzes ausgedrückt worden ist. Da dies nicht immer bis zu sogar in den wenigsten Fällen zutrifft, ist Informationsverlust eine zwangsläufige Folge. Die Schwächen des natürlichsprachlich gestützten Linkings wurden frühzeitig erkannt, und entsprechend zurückhaltend ist dieses Hilfsmittel bis jetzt auch benutzt worden.[81]

Hingegen wird weit verbreitet von den **Proximity-** oder **Adjacency-Operatoren** bei der Volltextrecherche (siehe Abschnitt 2.5.2.1.1) Gebrauch gemacht. Auch hier wird in der Fragestellung eine mehr oder weniger enge Nachbarschaft der Suchbegriffe in einem natursprachigen Text verlangt. Man macht sich also ebenso von den Zufälligkeiten der natürlichsprachigen Satzformulierung abhängig. Deswegen treten hier die gleichen Fehler auf wie beim Linking.[82]

2.4.3.5 Topologische Verfahren

Bei den **topologischen** oder **grafischen Verfahren** zur Darstellung von Begriffsverknüpfungen macht man von den Gesetzmässigkeiten Gebrauch, welche formal in der Verknüpfung von Punkten beliebiger Art herrschen. Diese Gesetzmässigkeiten werden in der Topologie untersucht, einem Teilgebiet der Mathematik.[83]

Die Elemente eines Gebildes beliebiger Art können in sehr verwickelter Weise durch Pfade miteinander verknüpft sein, welche ihrerseits wieder bestimmte Eigenschaften aufweisen, wie Kreuzung, Berührung, Ringbildung usw. Bei diesen Pfaden kann es sich beispielsweise um die Stationen eines Verkehrsnetzes handeln oder auch um die wechselseitigen Verknüpfungen der Atome innerhalb eines chemischen Moleküls. Auch die Verknüpfungen der Begriffe eines Textes können als geometrische Figur aufgefasst und mit topologischen Methoden abgebildet werden. So kann man ein exaktes Muster von allen engeren und weiteren Begriffsverknüpfungen aufbauen, an denen man interessiert ist. Beim Suchen in solchen topologischen Verknüpfungsmustern formuliert man die Begriffe der Fragestellung als Suchbedingungen. Sie bilden die Punkte im Verknüpfungsmuster. Zugleich fordert man auch die gewünschte Art ihrer syntaktisch-topologischen Verknüpfung.

Die wohl grösste Anwendung haben topologisch-syntaktische Methoden auf dem Gebiet der Chemie gefunden. Hier wird ein Molekül als ein Gebilde von Punkten und Verbindungslinien

[81] Fugmann 1999:91
[82] Fugmann 1999:58
[83] Fugmann 1999:92

aufgefasst, wobei die Punkte die Atome sind, aus denen das Molekül aufgebaut ist. Die Verbindungslinien werden von den chemischen Bindungen gebildet, die zwischen diesen Atomen bestehen. Fast 20 Millionen chemische Verbindungen sind bereits auf diese Weise gespeichert worden und stehen weltweit zur Recherche zur Verfügung. Jedes Jahr kommt eine halbe Million neuer chemischer Verbindungen dazu. Die Recherchenergebnisse erreichen bei diesem Verfahren fast 100% Genauigkeit und 100% Vollständigkeit, sofern die Suchbegriffe gut definiert sind (was bei chemischen Elementen der Fall ist). Allerdings ist das Verfahren sehr kostspielig.[84]

Trotz all dieser dargestellten syntaktischen Hilfsmittel für eine strukturierte Indexierung ist es laut Knorz offensichtlich, dass das Problem einer präzisen Benennung von Begriffen sowohl auf syntaktischem wie auch lexikalischem Niveau gelöst werden muss.[85] Je komplexer die Indexterme und je strukturierter die Indexierung, desto schwieriger und aufwändiger werden aber der Indexierungsprozess und auch das Retrieval. Dennoch könnte gerade die wachsende Bedeutung strukturierter (speziell multimedialer) Dokumente zum verstärkten Einsatz strukturierender bzw. syntaktischer Indexierungssprachen führen.[86] Die Probleme allerdings, welche im Zusammenhang mit einer komplexen Indexsprache auftauchen können, äussern sich in einem fast unvermeidbaren Aktualitäts- und Spezifitätsmangel der Indexsprache und einem enormen Erarbeitungs- und Erprobungsaufwand. Bei der konkreten Erstellung eines Index sollten die besprochenen Möglichkeiten und Entwurfsentscheidungen im Voraus festgelegt werden.

Damit schliesse ich die Erläuterungen zu den Indexsprachen ab, um nun die zahlreichen Verfahrensweisen des Indexierens zu demonstrieren, deren Unterscheidung natürlich auch mit der Wahl der Indexsprache zusammenhängt.

2.5 Indexierungsverfahren

Es gibt diverse Verfahren, um einen Index eines bestimmten Typs zu erstellen. In einem ersten Schritt kann man zwischen manuellen und automatischen Verfahren differenzieren. In einem zweiten Schritt werden Extraktionsverfahren und Zuteilungsverfahren voneinander abgegrenzt. Diese Verfahrenstechniken werden in den nachfolgenden Abschnitten vorgestellt.

Gelegentlich wird auch eine tiefe Indexierung von einer breiten unterschieden. **Tiefes Indexieren** bedeutet ausgesprochen erschöpfendes und spezifisches Indexieren und somit ein gutes Retrievalergebnis. Unter **breitem Indexieren** wird eine Indexierung mit wenigen, allgemeinen

[84] Fugmann 1999:93
[85] Knorz 1997:127
[86] Knorz 1997:127

Begriffen verstanden, wobei das Retrievalergebnis zwar weniger gut ist als beim tiefen Indexieren, dafür aber meist schneller und wirtschaftlicher durchzuführen ist.[87]

2.5.1 Manuelle Verfahren

Beim manuellen Indexieren weist ein professioneller Indexierer einem Dokument die Indexterme zu, manchmal ist es auch der Verfasser eines Textes selbst. Speziell bei der Erstellung von Buchregistern und Indexen für einzelne Texte kommen die intellektuellen Indexierungstechniken zum Einsatz. Innerhalb von grossen und heterogenen Dokumentkollektionen, wie z.B. im Internet, werden dagegen die automatischen Verfahren bevorzugt, denn eine manuelle Indexierung wäre dort gar nicht realisierbar. Der Vorteil der manuellen Indexierung liegt darin, dass ein gut ausgebildeter und erfahrener Indexierer den Inhalt eines Dokuments versteht und deshalb wirklich inhaltsbeschreibende Indexterme vergeben kann. Bei den manuellen Methoden (und auch bei den maschinellen Verfahren, die in Abschnitt 2.5.2 behandelt werden) trennt man zwischen Extraktions- und Zuteilungsmethoden bzw. zwischen **dokumentorientierten** und **begriffsorientierten Ansätzen**.[88]

2.5.1.1 Extraktionsmethoden

Eine Indexierung kann in der direkten Entnehmung der Indexterme aus dem Originaltext bestehen. Diese Prozedur wird **Extraktionsindexierung** (*extraction indexing*, *derivative indexing*) genannt, manchmal als Indexierung mit unkontrolliertem Vokabular bezeichnet. Die einem Text entnommenen Indexterme sind einzelne Wörter oder ganze Phrasen, die als Schlagwörter eingespeichert oder z.B. ins Register eines Buches übernommen werden. Die Anzahl der extrahierten Indexterme variiert von einigen wenigen bis zu sehr vielen, abhängig vom Bedarf, einen Textinhalt mehr oder weniger detailliert zu repräsentieren. Zur Vermeidung allzu grosser Heterogenität auf der sprachlichen Ebene werden in der Regel einige Vereinheitlichungen vorgenommen, welche die Form der Indexterme betreffen. Die Extraktionstechnik ist im Prinzip das einfachste Indexierungsverfahren.[89]

Eine Indexierung mit extrahierten natürlichsprachlichen Indextermen hat Vor- und Nachteile. Zuerst einmal besteht der Vorteil, dass die Indexierung sehr ausdrucksstark und flexibel ist. Ferner repräsentiert sie eine ganze Reihe von Zugangsmöglichkeiten und Perspektiven eines Texts, und es ist äusserst einfach, neue und komplexe Konzepte darzustellen. Das Indexvokabular ist weniger

[87] Kaiser 1996:3
[88] Kaiser 1996:6
[89] Knorz 1997:129; Moens 2000:51

streng kontrolliert bis überhaupt nicht, und somit kann eine grosse Vielfalt von Indexdeskriptoren oder Schlagwörtern identifiziert werden.[90] Da fixe Indexterme fehlen, besteht die Möglichkeit, aus den natürlichsprachlichen Indextermen eine portable und kompatible Textdatenbank für verschiedene Dokument- oder Textkollektionen herzustellen.

Natürlich haften der Methode auch Nachteile an. Die Textwörter haben die Eigenschaft, mehrdeutig zu sein, da jedes inhaltstragende Wort in einer Phrase den Kontext für die anderen Wörter darin bildet. Ausserdem sind die Wörter und Phrasen in einem Text häufig zu spezifisch, um einen Text zu repräsentieren. Sie verhindern so die Suche nach Allgemeinbegriffen und allgemeinen Informationskonzepten im Text. Es erweist sich als schwierig, die zugrunde liegenden Konzepte mit extrahierten Termen einzufangen.[91] Auch wenn ein lexikalischer Ausdruck für einen Begriff existieren würde, werden beim Extraktionsindexieren Umschreibungen dieses Begriffs nicht übersetzt und eine Recherche nach dem Begriff verunmöglicht. Es unterbleiben auch die Bedeutungserkennung, die Begriffsanalyse und das Auffüllen von Auslassungen. Das Verfahren ist nur dann zuverlässig in Bezug auf Recherchengenauigkeit, wenn für gesuchte Begriffe lexikalische natürlichsprachliche Ausdrücke existieren, die in einem zu indexierenden Text auch genannt werden. Nach Fugmann entsteht beim manuellen Extraktionsindexieren letztlich gar kein eigentlicher Index, sondern eine Konkordanz, das heisst eine Zusammenstellung von Fundstellen für Textwörter.[92]

2.5.1.1.1 Manuelle Stichwortindexierung

Manuelle Stichwortextraktion tritt meistens nur in speziellen Anwendungsbereichen auf. Sie kann während einer Texterfassung als zusätzliche Markierung von Textwörtern etwa zur Aufnahme in ein Register erledigt werden und als Option eines Textverarbeitungssystems angeboten werden. Das Indexierungsverfahren enthält eine tatsächliche dokumentbezogene Relevanzentscheidung.[93]

2.5.1.2 Zuteilungsmethoden

Eine Datenbank oder ein Register, das durch blosse Extraktion von Wörtern aus einem zu indexierenden Text entstanden ist, kann mit erheblichen Mängeln behaftet sein. Diese betreffen vor allem die Präzision sowie die Ambiguität der Indexterme. Überdies erfordern grafische

[90] Das vollständige Fehlen einer Vokabularkontrolle ist rar. Verschiedene morphologische Varianten eines Terms oder Synonyme eines Begriffs werden üblicherweise durch eine Standardform ersetzt. Beispielsweise wird die morphologisch am wenigsten besondere Form, meistens das Substantiv, für die selektierten Terme benutzt.
[91] Moens 2000:51
[92] Fugmann 1999:114
[93] Knorz 1997:129

Informationen in Gestalt von Bildern oder Tabellen die Zuteilung von inhaltsdarstellenden Indextermen.[94] Im Regelfall werden bei einer manuell vollzogenen Indexierung deshalb den Texten zum Zweck ihrer inhaltlichen Erschliessung inhaltskennzeichnende Wörter zugeteilt, Wörter also, die im Text nicht verbal vorkommen müssen. Die Methode bezeichnet man als **zuteilende Indexierung** (*assignment indexing*).[95]

Je nachdem, woher die zuzuweisenden Indexterme stammen und in welcher Zuverlässigkeit und Sorgfalt die Zuteilung erfolgt, unterscheidet man beim Zuteilungsindexieren zwischen dem **freiem Indexieren**, dem **kontrollierten Indexieren**, dem **klassifizierenden Indexieren** und dem **verbindlichen Indexieren**. Daneben gibt es auch Mischformen, die als **Hybrid-Indexierung** bezeichnet werden. Dem kontrollierten, klassifizierenden und verbindlichen Indexieren ist gemeinsam, dass sie für den Indexierungsprozess kontrollierte Vokabulare (wie in Abschnitt 2.4.2.2 erläutert) zu Hilfe nehmen.

2.5.1.2.1 Freies Indexieren

Beim freien Indexieren verfügt ein Indexierer über alle Freiheiten, treffend erscheinende Indexterme selbst zu prägen und den Texten hinzuzufügen.[96] Was die Wiedergabe der Essenz eines Textes anbetrifft, so verfügt der Indexierer hier im Gegensatz zum Extraktionsindexieren über die Möglichkeit, Paraphrasierungen, die er in den Texten antrifft, aufzulösen und in lexikalische Ausdrücke zu übersetzen. Es besteht jedoch keine Verpflichtung.

Das Plus des freien Indexierens liegt hauptsächlich darin, dass das Verfahren für einen Indexierer einfach ist, dass Indexterme sehr spezifisch erfasst und dass in Texten vorkommende Auslassungen oder Ellipsen aufgefüllt werden können. Fehlt es einem kontrollierten Indexvokabular an einer wünschenswerten Spezifität und somit an passenden Begriffen, kann dagegen beim freien Indexieren ein Indexterm frei formuliert und dem Spezifitätsbedürfnis der Indexierung entsprechend angepasst werden. Besteht die Aufgabe einer Indexierung ausschliesslich in der Erfassung von Individualbegriffen, für welche eine grosse terminologische Einheitlichkeit charakteristisch ist, so stellt das freie Indexieren eine gute Verfahrensoption dar. Ausserdem kann die Zeit, welche für die Suche nach einem Indexterm in einem Thesaurus oder in einer Klassifikation gebraucht wird, mit diesem Verfahren eingespart werden.

Sämtliche Nachteile eines unkontrollierten oder freien Indexvokabulars gelten auch automatisch als Negativpunkte für das freie Indexieren. Die Voraussehbarkeit der inhaltlichen Erschliessung und der Essenzwiedergabe ist eher gering, was bei einer Recherche zu Informationsverlust führen kann,

[94] Fugmann 1999:115
[95] Cleveland/Cleveland 1990:87; Fugmann 1999:115; Lancaster 1998:14; Moens 2000:51
[96] Fugmann 1999:19

denn die Begriffsbenennungen eines Indexierers und eines Benutzers können weit auseinander liegen. Die schlechte Voraussehbarkeit ergibt sich auch daraus, dass ein Indexierer sich nicht nur an keinen vorgegebenen Wortschatz halten muss, sondern dass er sich auch nicht nach irgendeinem Schema an begrifflichen Kategorien richten muss. Was den Zeitgewinn anbelangt, den man sich vom freien Indexieren verspricht, so kann er sich leicht als Trugschluss erweisen. Hat ein Indexierer nämlich kein Vokabular zur Verfügung, welches ihm die entsprechenden Vorschläge für Begriffsbenennungen unterbreitet, dann kann das Kopfzerbrechen lange dauern, bis er einen treffenden Indexterm gefunden hat und sich damit zufrieden gibt. Die grosse Last liegt bei der freien Indexierung auf Seiten des Fragestellers. Mit Raten und Probieren muss er versuchen, die richtigen Suchterme herauszufinden, insbesondere bei der Recherche nach Allgemeinbegriffen.

Freies Indexieren leistet nützliche Dienste beim Sammeln von Fachausdrücken und kann so die Grundlage bilden für den Aufbau eines Indexsprachenwortschatzes, also z.B. eines Thesaurus. Schon während des Stadiums des Sammelns kann man einen ersten Überblick darüber gewinnen, welche Begriffe man im Vokabular in welcher Spezifität benötigen wird oder mit welchen Kategorien zu rechnen ist, und man sammelt Erfahrungen, wie man den späteren Wortschatz gliedern und unterteilen sollte.[97]

2.5.1.2.2 Kontrolliertes Indexieren

Beim kontrollierten Indexieren (*controlled indexing*) werden ausschliesslich die im Indexsprachenwortschatz enthaltenen Begriffsbenennungen für die Indexierung verwendet. Es handelt sich beim Wortschatz um ein kontrolliertes Vokabular, um eine Zusammenstellung von Begriffsbenennungen, die für das Indexieren zugelassen sind. Wie wir bereits gesehen haben, sind Thesauri, Klassifikationen und Schlagwortlisten die üblichen Werkzeuge.[98]

Zwar wird mit dem kontrollierten Indexieren die unerwünschte Vielfältigkeit der natürlichsprachlichen Ausdrucksweise eingeschränkt und die Voraussehbarkeit der gespeicherten Ausdrucksweisen verbessert. Allerdings wird dabei manchmal übersehen, dass der Wortschatz nicht nur dem Indexierer gehört, sondern auch dem Fragesteller. Der Indexierer darf deswegen diesen Wortschatz immer nur so verwenden, dass er auch vom Benutzer nachvollzogen werden kann. Geschieht dies nicht, können Ballast und Informationsverlust die Folge sein.

Die Vorteile des kontrollierten Indexierens liegen in der Allgemeinheit, der Nichtmehrdeutigkeit und der Genauigkeit der Indexterme. Die Suche nach Allgemeinbegriffen, und natürlich auch nach Individualbegriffen, ist möglich. Die nicht mehrdeutigen Indexterme kontrollieren die Vielfältigkeiten der sprachlichen Ausdrucksweisen von identischen oder ähnlichen Konzepten und

[97] Fugmann 1999:120
[98] Fugmann 1999:115

befassen sich dadurch mit Synonymie-, Hierarchie- und Assoziationsrelationen zwischen Termen. Sie können auch leicht in andere Sprachen übersetzt werden und in mehrsprachigen Informationssystemen angewendet werden.

Aber das kontrollierte Indexieren erlaubt nur wenige Zugangsmöglichkeiten zu einem Text und wenige Repräsentationsperspektiven. Die Indexierung wird unflexibel und nimmt keine Rücksicht auf die Bedürfnisse der Textbenutzer.[99] Hinzu kommen alle Nachteile, die bereits beim kontrollierten Vokabular aufgezählt wurden, also der sorgfältige Aufbau und die Pflege des Vokabulars, die höheren Kosten und der grössere Zeitaufwand des Indexierungsprozesses, die stark gefährdete Aktualität und Spezifität des Vokabulars sowie die eingeschränkte Freiheit des Indexierers. Sobald verschiedene kontrollierte Vokabulare nicht austauschbar sind, sind auch die darauf basierenden Retrievalsysteme weniger portabel und kompatibel zu anderen Dokumentkollektionen.

Steht kein Thesaurus oder Schlagwortregister von vornherein zur Verfügung, so kann man dennoch besonders zuverlässig indexieren, wenn man gleichzeitig mit dem Indexieren einen Thesaurus oder ein Schlagwortverzeichnis aufbaut. Dann erhält man einen guten Überblick über die nahe verwandten Schlagwörter oder Deskriptoren, was sehr zur Vollständigkeit des gesamten Netzwerkes der Verweisungen beiträgt. Diese Arbeitsweise ist für das Indexieren von Büchern erst neuerdings vereinzelt empfohlen worden.[100]

2.5.1.2.3 Klassifizierendes Indexieren

Unter klassifizierendem Indexieren wird eine besondere Art des kontrollierten Indexierens verstanden: das Zuteilen von Notationen aus einem Klassifikationssystem zu den inhaltlich zu erschliessenden Dokumenten.[101] Dies bedeutet, dass unter kontrolliertem Indexieren die Anwendung von Thesauri, Classauri oder Schlagwortlisten gemeint ist, von klassifizierendem Indexieren die Rede ist, sobald eine Klassifikation zur Anwendung kommt.

Klassifizierende Indexierung beruht auf der Annahme, dass die Indexierung mittels kontrolliertem Vokabular mit der Textklassifikation verwandt ist. **Textklassifikation** (*text classification*) meint die Gruppierung von Texten, die konzeptuell in enger Beziehung stehen, zu Textklassen.[102]

Das klassifizierende Indexieren profitiert und leidet unter sämtlichen Vor- und Nachteilen des kontrollierten Indexierungsverfahren sowie der Vokabularkontrolle. Sie werden hier nicht noch einmal aufgezählt.

---

[99] Moens 2000:53
[100] Mulvany 1994:278
[101] Fugman 1999:119; Moens 2000:53
[102] Lancaster 1998:16; Moens 2000:53

2.5.1.2.4 Verbindliches Indexieren

Zur Vermeidung einiger Mängel des kontrollierten Indexierens kann der Weg zum verbindlichen Indexieren eingeschlagen werden. Es handelt sich dabei um eine Indexierungsart, bei welcher nur die jeweils bestpassenden Deskriptoren, Schlagwörter oder Notationen aus einem Indexvokabular benutzt werden dürfen.[103] Das Ziel dabei ist, dass nicht nur der Fragesteller, sondern auch der Indexierer sich auf die Suche nach den jeweils bestpassenden Begriffsbenennungen im gemeinsamen Wortschatz begeben muss. Nur so wird die Voraussehbarkeit der Ausdrucksweisen für die Indexterme gewährleistet, und nur so kann ein Benutzer erahnen, welches Schlagwort oder welchen Deskriptor ein Indexierer für den Begriff seines Interesses gewählt haben könnte.

Diese Arbeitsweise ist schon gegen Ende des vergangenen Jahrhunderts in Amerika von Cutter im Bibliothekswesen als erforderlich beschrieben worden. Sie ist seitdem als Selbstverständlichkeit auch bei jeglichem Klassifizieren in den Bibliotheken befolgt worden, beim modernen Indexieren mit Thesauri und Klassifikationen jedoch etwas in Vergessenheit geraten.[104]

Man könnte einwenden, dass die Verwendung der bestpassenden Indexterme doch eigentlich bereits beim kontrollierten Indexieren angestrebt wird und die Unterteilung in kontrolliertes und verbindliches Indexieren gar nicht notwendig ist. Dennoch scheint es wichtig, zu erwähnen, dass die kontrollierte Indexierung gegenüber einer freien bloss überlegen sein kann, wenn die Indexierung auch verbindlich ist. Ausschliesslich unter den Bedingungen des verbindlichen Indexierens kann der kontrollierte Wortschatz in vollem Umfang seinen Zweck erfüllen und eine möglichst verlust- und ballastfreie Recherche realisiert werden. Eine kontrollierte Indexierung allein ist also noch keine Voraussetzung für eine gute Indexierung.

Aus diesen Überlegungen heraus ergibt sich die Notwendigkeit, dass ein Indexsprachenwortschatz möglichst übersichtlich gestaltet sein sollte, um das schnelle und sichere Gelangen zu den jeweils bestpassenden Schlagwörtern und Deskriptoren immer zu gewährleisten, insbesondere auch unter Zeitdruck. Die Übersichtlichkeit eines Vokabulars und eine Ordnung darin können erreicht werden, indem der Wortschatz kategorisiert, hierarchisch gegliedert, logische und natürlich nachvollziehbare Unterteilungsgesichtspunkte angewendet, Präkombinationen vermieden und eine Indexsprachengrammatik eingeführt werden.

[103] Fugmann 1999:118
[104] Fugmann 1999:118

2.5.1.2.5 Hybrid-Indexieren (manuell zuteilend)

Durch die Kombination von kontrollierter und freier Indexsprache ist es möglich, die spezifischen Schwächen beider Vokabulararten zu überwinden und ihre jeweiligen Stärken zu nutzen. Insbesondere kann freies Indexieren eine wertvolle Ergänzung zum Indexieren mit vorgegebenem Vokabular bieten, weil es die bei den Indexsprachen fast unvermeidliche Aktualitäts- und Spezifitätslücke wenigstens teilweise zu schliessen mag.[105]

Das aus der Kombination resultierende Indexierungsverfahren ist das manuelle Hybrid-Indexieren. Es arbeitet sozusagen mit einem zweigeteilten Wortschatz. Der Kern des Wortschatzes besteht aus den möglichst verbindlich benutzten Begriffsbenennungen. Das Randvokabular besteht aus den frei gewählten zusätzlichen Begriffsbenennungen. Von Zeit zu Zeit müssen Entscheidungen darüber gemacht werden, ob bei gewissen Begriffsbenennungen der Bedarf für eine Aufnahme ins Kernvokabular besteht. So können viele Nachteile des verbindlichen und des freien Indexierens getilgt werden, z.B. die durch das Randvokabular ermöglichte Zuweisung von Neologismen. Allerdings leidet das Randvokabular zwangsläufig unter den Schwächen des freien Indexwortschatzes und das Kernvokabular unter denjenigen des verbindlichen.

Extraktionsindexieren ist für den Zweck einer Kernvokabularergänzung gemäss Fugmann weniger geeignet, weil es nicht die Lexikalisierung der Paraphrasierungen zulässt und auch nicht das Auffüllen von Ellipsen. Es unterbleiben ebenso die Bedeutungsklärung bei mehrdeutigen Wörtern und die Begriffsanalyse.[106] Der Ansatz müsste für eine aussagekräftige Beurteilung allerdings zuerst sorgfältig getestet werden.

2.5.2 Automatische Verfahren

Schon seit den 50er Jahren haben die Fachleute der Informations- und Dokumentationsbranche versucht, den Indexierungsprozess maschinell durchzuführen. Was die Computerkapazitäten und die Umsetzung von gedruckten Texten in maschinenlesbare Formen angeht, existieren heutzutage eigentlich keine Probleme mehr. Die Übersetzung der Indexierungsregeln in einen Algorithmus und dessen exakte Formulierung aber konnten bis heute nicht wirklich gelöst werden. Doch in den letzten Jahren ist die Informationsflut stark angestiegen, und auch in absehbarer Zeit wird die Menge der Informationen und der in natürlicher Sprache verfassten Texte und elektronischen Dokumente weiter anwachsen. Diese Umstände lassen das Interesse am Management von Informationen und an einer automatischen Indexierung immer wieder aufleben, wenn nötig auch mit Techniken, die eigentlich nichts mit einer intellektuellen Indexierung gemeinsam haben. Die gängigen Verfahren zur automatischen Indexierung basieren vor allem auf der Analyse der

[105] Fugmann 1999:121; Knorz 1997:125
[106] Fugmann 1999:122

Worthäufigkeit (statistische und probabilistische Verfahren) und haben den Vorteil der grossen Zeitersparnis beim Indexierungsvorgang.

Automatische Indexierung (*automatic indexing*) hat zum Ziel, Informationen mittels Computer zu repräsentieren, damit sie zu einem späteren Zeitpunkt wiederauffindbar sind. Der Prozess beinhaltet, wie auch die manuelle Indexierung, die Inhaltsanalyse eines Textes, Selektion und Generalisierung der Information sowie die Übersetzung in eine Endform.

Von **maschinell unterstützter** oder **computerunterstützter Indexierung** (*computer-assisted indexing*, *computer-aided indexing*) spricht man, wenn ein automatisches Verfahren Indexterme vorschlägt, die in einer manuellen Nachbearbeitung noch bestätigt werden müssen. Der Unterschied zur automatischen Indexierung besteht darin, dass Computer zur Verwendung von Routinearbeiten genutzt werden, während der Mensch die intellektuellen Aufgaben löst.[107]

Für die computerunterstützte Indexierung wird von Cleveland/Cleveland eine Drei-Schritt-Prozedur vorgeschlagen:

1. Ein Indexierer überfliegt ein Dokument und wählt die Stellen aus, die für das Indexieren in Frage kommen (z.B. Titel, Kapitelüberschriften, Sätze, die den Zweck, die Methoden und Resultate beschreiben, usw.). Dieser Schritt kann auch von unerfahrenen oder nichtprofessionellen Indexierern ausgeführt werden.

2. Das zum Indexieren verwendete Material wird von einem Computer eingelesen, und dieser produziert mit einem der Standardextraktionsverfahren sowie Termgewichtung (siehe Abschnitt 2.5.2.1.3.1) potenzielle Indexterme.

3. Ein professioneller Indexierer editiert das Resultat des Computers und unternimmt alle als notwendig betrachteten Änderungen.[108]

Heutzutage ist manuelles Indexieren eigentlich immer computerunterstützt. Es ermöglicht, die Vorteile der automatischen und der manuellen Indexierung zu verbinden. In den folgenden Abschnitten werden wir uns mit den Verfahren des automatischen Indexierens beschäftigen. Die Gliederung in Extraktions- und Zuteilungsmethoden der manuellen Indexierungsverfahren wird beibehalten.

[107] Cleveland/Cleveland 1990:228; Kaiser 1996:38f.; Knorz 1997:128
[108] Cleveland/Cleveland 1999:228f.

2.5.2.1 Extraktionsmethoden

Der Umstand, dass eigentlich alle sinntragenden Wörter in einem Titel oder in einer Kurzfassung eines Textes recherchierbar sind, kann als eine vom Verfahren her „triviale, wenngleich zweifellos effektive Art von automatischer Indexierung“ aufgefasst werden.[109] Für maschinelle Extraktionsverfahren werden sehr häufig nur Titel und Kurzfassung (auch Referat oder Abstract genannt)[110] als textuelle Basis genommen und auf eine Bearbeitung des gesamten Originaldokuments verzichtet. Die Indexierung hängt dabei stark von der Qualität des Abstracts ab. Ebenso hängt die Indexierung des vollständigen Originaltextes von der Wortwahl des Verfassers ab, wie bei den manuellen Extraktionsmethoden schon erwähnt, und das maschinelle Verfahren weist auch dieselben Probleme auf wie das manuelle. Der grösste Vorteil des automatischen Extrahierens liegt in der Schnelligkeit, Einfachheit und Kostengünstigkeit des Verfahrens.

Wir unterscheiden innerhalb der automatischen Extraktionsmethoden die Volltextspeicherung, linguistische Verfahren sowie die kollektions- und retrievalorientierten Verfahren.

2.5.2.1.1 Volltextspeicherung

Aus Sicht der Datenbankproduzenten ist die **Volltextspeicherung** (*full-text search*) bzw. das **Freitextverfahren** oder die **Volltextinvertierung** die ideale Erschliessungsmethode. Denn die Speicherung aller in einem Text vorkommenden Wortformen ausser den Funktionswörtern ist schnell und billig. Funktionswörter sind sprachliche Elemente, die primär grammatische anstelle von lexikalischer Bedeutung tragen und vor allem syntaktisch-strukturelle Funktionen erfüllen (z.B. Artikel, Pronomen, Konjunktionen, Präpositionen).

Bei einer Volltextspeicherung wird eine Datei angelegt, in der alle Indexterme mit einem Verweis auf die Dokumente, in denen der Term vorkommt, enthalten sind. Der Indexierungsvorgang läuft in zwei Schritten ab:

1. Durch Extraktion aller Begriffe aus den Dokumenten mit dem entsprechenden Verweis auf das Dokument, aus dem der Begriff genommen wurde, wird eine invertierte Datei erzeugt.

2. Alle Funktionswörter werden aus der Indexdatei eliminiert. [111]

Die Basistechnik beim Retrieval im Zusammenhang mit der Volltextspeicherung ist das Bool’sche Retrieval (Erläuterungen dazu folgen in Abschnitt 2.5.2.1.3.1).

[109] Knorz 1997:121
[110] Kuhlen 1997:88
[111] Kaiser 1996:7

Das Verfahren hat deutliche Vorteile: Es ist völlig unabhängig von der Art der zu indexierenden Dokumente und deren Sprache. Die manuelle Pflege eines Indexvokabulars wird nicht erfordert, Massenspeicher sind billig geworden, Arbeitskräfte können eingespart werden, und Indexierungsfehler werden vermieden. Die Technik ist von einem wirtschaftlichen Standpunkt aus sehr attraktiv.[112] Ausserdem gibt es Fälle, wo Dokumente mit zugeteilten Deskriptoren oder Schlagwörtern nicht gefunden werden, weil der Informationsgehalt des Dokuments für den Hauptfokus eines Benutzers nur von peripherer Bedeutung ist. Gerade für unerfahrene Fragesteller scheint es einfacher zu sein, mit natürlichsprachlichen Ausdrücken in einem Volltextspeicher zu suchen als eine Suche mit fixen Indextermen.

Die nichtinterpretierte Speicherung von Volltexten hat auch einen erheblichen Mangel. Es werden sehr oft unerwünschterweise Dinge gefunden, die nichts mit dem Suchbegriff zu tun haben. So werden bei einer Suche nach „Silber“ beispielsweise auch die „Silbertanne“ oder die „Silberne Hochzeit“ zurückgegeben. Andererseits findet man aber auch vieles, was damit zusammenhängt. Für das Retrieval von Individualbegriffen, wie z.B. für Personen, Institutionen, Städte usw., leistet das Extraktionsindexieren gute Dienste. Hat man es sogar ausschliesslich mit Individualbegriffen zu tun (Telefonverzeichnis, Autorenregister), dann kommt man gut mit der Volltextspeicherung aus. Viel schwieriger erweist sich die Suche nach Allgemeinbegriffen, da diese in den Texten zumeist in unvorhersehbarer Form ausgedrückt sind oder überhaupt nicht. Auch Begriffsverknüpfungen wie Synonymie- oder Hierarchiebeziehungen bleiben unterlassen.

Die blosse Suche nach Textwörtern in einem Volltextspeicher kann zu Informationsverlust und Ballast führen. Der Erschliessungsaufwand wird auf den Fragesteller verlagert. Die Freitextverfahren können jedoch den kostspieligen und zeitraubenden Beschaffungsweg von Dokumenten drastisch abkürzen.[113]

2.5.2.1.2 Linguistische Verfahren

Um der Tatsache, dass die Volltextspeicherung eigentlich keine wirkliche Inhaltserschliessung anbietet, entgegenzuwirken, wurden Verfahren entwickelt, welche die sprachlichen Probleme der Freitextverfahren mit linguistischen Methoden angehen. Die meisten dieser linguistisch motivierten Verfahren selektieren aus einem zu indexierenden Text die natürlichsprachlichen Wortformen, die dann als Indexterme verwendet werden. Diese Indexterme sollen den Inhalt eines Dokuments reflektieren und können aus einzelnen Wörtern oder aus Mehrwortausdrücken bzw. Phrasen bestehen. Es war Luhn, der 1957 erstmals den Vorschlag machte, dass bestimmte Wörter aus einem

[112] Moens 2000:17
[113] Fugmann 1999:194

Text extrahiert werden sollten, um dessen Inhalt zu repräsentieren.[114] Auch heute noch werden Dokumente nach diesem Prinzip indexiert. Allerdings sind nicht alle Wörter in einem Text gute Indextermkandidaten, und nicht alle guten Indextermkandidaten definieren den Textinhalt gleich gut. Es gibt deshalb verschiedene Techniken, welche die inhaltstragenden Indexterme identifizieren sollen.

Bei einer Extraktion von natürlichsprachlichen Indextermen mit linguistischen Methoden werden gemäss Salton die folgenden Schritte unternommen (die Reihenfolge der Schritte kann variieren):

1. Mit einer lexikalischen Analyse werden die einzelnen Wörter in einem Text identifiziert.

2. Funktionswörter und in einem Fachgebiet sehr häufig auftretende Wörter, die nicht genügend spezifisch sind, um den Inhalt eines Textes angemessen darzustellen, werden mithilfe einer Stoppwortliste entfernt.

3. Optional werden die übrig bleibenden Wörter auf ihre Stammform reduziert (Stemming).

4. Optional werden die Stammformwörter zu Phrasen formiert, die als Indexterme dienen.[115]

Die lexikalische Analyse beginnt, wenn ein Text bereits elektronisch gespeichert ist. Die einzelnen **Tokens** (damit sind alle einzelnen Vorkommen von Wortformen gemeint) werden identifiziert und daraus die Indextermkandidaten produziert. In Texten auftretende Zahlen sind keine guten Indexterme; sie werden meistens ignoriert.[116]

Beim zweiten Schritt werden die Indextermkandidaten mit **Stoppwortlisten** (*stoplist*) oder **Negativlisten** (*negative dictionary*) gefiltert und die nichtbedeutungstragenden Wörter von der weiteren Verarbeitung ausgeschlossen. Die maschinenlesbaren Listen bestehen aus Wörtern, die nicht als Indexterme gewählt werden dürfen. Solche Listen differieren in ihrer Grösse und können von 50 bis 400 Wortformen enthalten. Bis zu 50% aller Tokens werden eingespart.[117] Je früher der Gebrauch der Stoppwortliste stattfindet, desto effizienter und speicherplatzsparender werden die nachfolgenden Verarbeitungsprozesse.

Stoppwortlisten werden vor dem eigentlichen Indexierungsprozess erstellt. Es gibt verschiedene Techniken dafür. Üblicherweise werden alle Funktionswörter aufgenommen. Denn ihre Verwendung als Indexterme ist kritisch. Die Wörter werden also aufgrund ihrer Wortkategorie selektiert.[118] Eine ebenfalls weit verbreitete Technik zum Aufbau einer Stoppwortliste ist die Aufnahme aller Wörter, die am häufigsten in Texten auftreten, entweder generell oder auf einen

[114] Moens 2000:77
[115] Salton 1989:303ff.
[116] Moens 2000:79
[117] Knorz 1997:130; Moens 2000:80
[118] Moens 2000:80

Fachbereich bezogen.[119] Die Festlegung von Grenz- oder Schwellwerten bestimmt die Inklusion oder Exklusion von Wörtern. Die Auftretenshäufigkeit eines Wortes ist aber kein hundertprozentiges Kriterium für dessen inhaltstragende Bedeutung. Weil Funktionswörter zu Kürze tendieren, werden gelegentlich alle Wörter, die kleiner sind als ein Schwellwert, in eine Stoppwortliste aufgenommen. Mit einer Anti-Stoppwortliste wird danach eine Eliminierung von wichtigen inhaltstragenden kurzen Wörtern verhindert. Eine etwas andere Methode verwendet eine Trainingskollektion von Texten sowie die Informationen über ihre Beziehungen zueinander. Ein Punktesystem reflektiert die Wichtigkeit der Wörter, indem Texte, die das gleiche Thema behandeln und miteinander verknüpft sind, erkannt werden. Das Punktesystem basiert auf der Wortverteilung in den verknüpften Texten. Als Stoppwörter werden dann die Wörter mit niedriger Punktzahl auserwählt.

Der dritte Schritt bei der sprachbasierten Extraktion von Indextermen ist die **Stemmatisierung** (*stemming*). Morphologische Varianten von Wortformen werden auf den Stamm zurückgeführt, weil angenommen wird, dass Wörter mit demselben Stamm semantisch verknüpft sind und für den Benutzer eines Textes die gleiche Bedeutung haben.[120] Im Umfeld des Information Retrieval soll mit der Wortformenreduzierung eine bessere Übereinstimmung zwischen den Indextermen eines Dokuments und den Termen einer Suchanfrage erzielt werden. Die Indexterme werden sozusagen „erweitert“.

Die automatische Stemmatisierung kann in einer **Überstemmatisierung** (*overstemming*) oder **Unterstemmatisierung** (*understemming*) resultieren. Für morphologiereiche Sprachen wie das Deutsche ist die Technik dennoch nützlich, ebenso für kurze Texte. Die Entfernung von Flexionsmorphemen wirkt sich nur geringfügig auf die Wortbedeutungen aus und kann gefahrlos durchgeführt werden; die Entfernung von Derivationsmorphemen dagegen kann die Wortbedeutungen verändern. Im Verlauf der Untersuchung der Extraktionsmethode mit der inversen Seitenhäufigkeit zur automatischen Indexierung von deutschsprachigen Texten im Abschnitt 5.3 werden die Flexionsmorpheme in Anlehnung an die hier gemachten Ausführungen entfernt werden, die Derivationsmorpheme jedoch erhalten bleiben. Es herrscht in der Fachliteratur des Indexierens Einigkeit darüber, dass eine Stemmatisierung, auch wenn sie in zahlreichen Fällen die Performanz eines Retrievalsystems nicht verbessern kann, zumindest keine negative Einwirkung darauf hat.[121]

Die analytische Behandlung von Komposita hat einen effektiven Einfluss auf die Verbesserung der Retrievalergebnisse, insbesondere bei Sprachen wie dem Deutschen oder dem Holländischen. Sie lässt sich wörterbuchabhängig recht einfach realisieren. Aufwändigere Verfahren arbeiten mit einer partiellen oder vollständigen syntaktischen Analyse. Bei der Verarbeitung der extrahierten Phrasen aus den untersuchten Buchtexten zur Erstellung von Schlagwortregistern werden die Komposita

[119] Salton 1989:279
[120] Moens 2000:81
[121] Knorz 1997:131; Moens 2000:81

nicht in ihre Bestandteile zerlegt, um die Erzeugung von Ballast zu vermeiden. In Abschnitt 5.2.2 werde ich nochmals auf die Kompositaanalyse im Zusammenhang mit automatischer Indexierung zu sprechen kommen.

Für den vierten Schritt, die Selektion der Phrasen, wird von der Annahme ausgegangen, dass die Problemfelder der Volltextrecherche sich nicht nur auf die Wortebene beschränken und dass Elemente der syntaktischen Ebene mehr semantische Bedeutung tragen als einzelne Wörter. Es sind vor allem die **Nominal- und Präpositionalphrasen** (*noun and prepositional phrases*), von welchen man glaubt, dass sie inhalts- oder informationstragende Einheiten und deshalb gute Indikatoren für den Inhalt eines Textes sind. Man kann eine Phrase als Spezifikation eines Konzepts oder eines Begriffs im Sinne unserer Definition von Abschnitt 2.4.2.1 betrachten. Phrasen erhöhen also die Spezifität einer Indexsprache und auch die Präzision einer Suche.[122] Die Gliederung einer Nominalphrase in **Kopf** (*head*) und **Modifikator** (*modifier*) lässt beispielsweise automatisch Begriffsrelationen aufstellen. Ausserdem sind Phrasen in Bezug auf die Bedeutung weniger ambig als die einzelnen Wörter, aus denen sie gebildet sind. Für die Erkennung von Phrasen sind grössere Computeranforderungen vorhanden (die Phrasen müssen erkannt und normalisiert werden), so sind dennoch Phrasen die natürlichsprachlichen Hauptkandidaten für Indexterme bei einer Textrepräsentation mittels Indexierung.[123] Obschon die Nachteile der syntaktischen Methoden generell in der Nachfrage nach hoher Computerkapazität, Speicherplatz und der Software, die zur Verfügung stehen muss, liegen, reicht für die Analyse von Nominal- und Präpositionalphrasen eine partielle Analyse aus. Die Phrasenindexierung vermag allerdings trotz allem die Retrievalergebnisse nur sehr schwach zu verbessern.

Wir beschäftigen uns im Folgenden kurz mit der Normalisierung von Phrasen, da dieser Prozess für die Gewichtung der extrahierten Phrasen im zweiten Teil der Arbeit von Bedeutung ist.

Wenn in einem Informationssystem Phrasen im Textindex und in den Suchanfragen auftauchen, so können zwei oder mehr Phrasen auf denselben Begriff referieren. Allerdings müssen die Phrasen dazu dieselbe Form haben, um miteinander abgeglichen werden zu können, was nur sehr selten der Fall ist. Ein Konzept oder ein komplexer Begriff kann in verschiedenen syntaktischen Strukturen ausgedrückt werden (z.B. „eine Gartenparty“, „eine Party im Garten“), mit lexikalischen Variationen oder morphologischen Varianten kombiniert sein, womöglich Anaphern und Ellipsen enthalten. Eine Normalisierung der Formulierungsvarianten und Paraphrasierungen von Phrasen ist bei der automatischen Indexierung deswegen obligatorisch.[124]

[122] Knorz 1994:14
[123] Godby 1994:1; Moens 2000:84; Salton 1989:294
[124] Moens 2000:87

Mögliche Methoden zur Normalisierung von Phrasen sind:

1. Eine einfache Methode benutzt ein maschinenlesbares Wörterbuch von Phrasenvarianten. Solche Wörterbücher sind üblicherweise fachgebietsspezifisch.

2. Die Entfernung von Funktionswörtern und eine Umstellung der Wortreihenfolge der übrig bleibenden Inhaltswörter ist ebenfalls eine einfache, jedoch nicht immer zuverlässige Methode der Phrasennormalisierung.

3. Eine sicherere Methode basiert auf der syntaktischen Phrasenidentifizierung mit einem Parser und der Definition von Metaregeln zur Erkennung von äquivalenten Phrasen. Sie kann mit Anaphernresolution und Wortstemmatisierung kombiniert werden.

Bei der Normalisierung der aus den Testdaten extrahierten Phrasen wird in dieser Arbeit die zweite Methode zur Anwendung kommen und dabei auch versucht, deren Nützlichkeit für eine automatische Indexierung herauszuarbeiten (siehe Abschnitt 5.3).

Ein Spezialfall von Phrasenerkennung stellt die Erkennung von Eigennamen dar. Sie gehören zu den Individualbegriffen und sind in vielen Retrievalanwendungen nützlich. Als Eigennamen gelten Personennamen, Firmennamen, Institutionen, Produktenamen, Orte, Währungen usw. Zwei Wege werden meistens eingeschlagen, um Eigennamen zu identifizieren:

1. Ein maschinenlesbares Lexikon oder Namenwörterbuch wird verwendet.

2. Weil viele Eigennamen neu auftauchen, verschwinden oder sich verändern, erfordert eine sorgfältige Identifizierung auch das Erkennen von neuen Namen. Häufig werden sie mit besonderen Regeln erkannt, welche die typischen Merkmale von Eigennamen erfassen, z.B. Grossschreibung, linguistischer Kontext, Indikatorwörter.

Die grosse Fluktuation der Eigennamen ist ein Problem für die erste Methode. Die vielen Varianten eines Namens und die zahlreichen Schreibweisen bereiten beiden Vorgehen Schwierigkeiten.

Bei der Auswertung und dem Vergleich der Text- und Indexdaten im statistischen Teil werden in Korrespondenz mit den hier gemachten Ausführungen die Nominal- und Präpositionalphrasen als Indextermkandidaten verwendet. Eigennamen erfahren keine besondere Behandlung, sie werden analog zu den „normalen“ Nominal- und Präpositionalphrasen behandelt.

Erst neuere linguistische Ansätze versuchen auch auf der Textebene einen Text automatisch aufzubereiten und für ein Retrieval nutzbar zu machen. Dabei geht es um die Zerlegung eines Textes in sinnvolle einzelne Bestandteile, deren Rollen und deren Beziehungen untereinander es in einem zweiten Schritt abzuklären gilt. Ein Forschungsgebiet, in welchem der Textebene eine

besondere Bedeutung zukommt, ist das des automatischen Abstracting – eine Aufgabe, die noch weitaus schwieriger ist als die automatische Indexierung.[125]

2.5.2.1.3 Statistische Verfahren

Ein vollständig anderer Weg zur automatischen Indexierung wird von der Information Retrieval-Forschung der letzten 30 Jahre vorgeschlagen. Im Mittelpunkt des Interesses stehen hier Systeme, die Fragerepräsentationen und Dokumentrepräsentationen in Form gewichteter Indexterme miteinander vergleichen und als Antwort eine Dokumentreihenfolge liefern. Sortierkriterium ist die vom System geschätzte Wahrscheinlichkeit bzw. Plausibilität von Relevanz auf eine gestellte Frage. Da das Retrievalverfahren nicht von einer manuellen Suchstrategie abhängt, sondern fest vorgegeben ist, besteht die Aufgabe darin, die Parameter des Verfahrens (Gewichtungen, Rechenvorschriften) zu optimieren.[126]

Termgewichtung, Vektormodelle und probabilistische Modelle sind die üblichen statistischen Verfahren des maschinellen Indexierens.[127] Sie werden häufig auch als kollektions- oder retrievalorientierte Verfahren bezeichnet.[128]

2.5.2.1.3.1 Termgewichtung

Ausgangspunkt der Überlegungen bei der **Termgewichtung** (*term weighting*) ist der einfache Vergleich der sowohl in der Suchfrage als auch im Dokument vorkommenden Indexterme (ermittelt z.B. durch ein einfaches Extraktionsverfahren). Es ist allerdings nahe liegend, dass die so ermittelte Trefferzahl nicht unbedingt ein besonders guter Indikator für Dokumentrelevanz sein muss. Eine Lösungsmöglichkeit besteht darin, die Indexterme geeignet zu gewichten. Automatisch zu indexieren heisst im Zusammenhang mit Termgewichtung also, Gewichtungen für die Wörter eines Dokuments zu berechnen. Zur Gewichtung von Indextermen gibt es verschiedene Möglichkeiten, die auf Luhns Modell der Worthäufigkeiten (1957) zurückgehen und die auf den folgenden zwei Grundannahmen basieren: [129]

- Je häufiger ein Wort in einem Dokument auftritt, desto wichtiger ist es für das Thema des Dokuments.

[125] Knorz 1994:15
[126] Knorz 1997:131-133
[127] Kaiser 1996:7-19; Knorz 1994:15-20
[128] Knorz 1997:133
[129] Aas/Eikvil 1999:5-8; Kaiser 1996:7-12; Knorz 1994:15f.; Larson/Hearst 1998:14-27; van Rijsbergen 1979:8-15

- Je häufiger ein Wort in allen Dokumenten innerhalb einer Kollektion auftritt, desto schlechter diskriminiert es die Dokumente voneinander.

Die einfachste Art der Termgewichtung ist die **Bool'sche Gewichtung** (*Boolean weighting*). Das Gewicht w eines Indexterms i ist 1, wenn das Wort in einem Dokument d vorkommt, und 0 im andern Fall. f_{id} ist die Frequenz des Wortes i im Dokument d.

$$w_{id} = \begin{cases} 1 \text{ wenn } f_{id} > 0 \\ 0 \text{ wenn } f_{id} = 0 \end{cases}$$

Die **Wortfrequenzgewichtung tf** (*word frequency weighting*) ist eine weitere einfache Art der Termgewichtung und bezieht sich auf die Auftretenshäufigkeit oder Frequenz f eines Terms i in einem Dokument d.

$$w_{id} = f_{id}$$

Die bisherigen Formeln zur Berechnung der Termgewichtung berücksichtigen nicht die Frequenz eines Terms über alle Dokumente hinweg in einer Kollektion. Ein weit verbreiteter Ansatz, der diese Information mit einbezieht, ist die **Inverse Dokumenthäufigkeit idf** (*inverse document frequency*). Die Basis für die inverse Dokumenthäufigkeit bildet die Beobachtung, dass Autoren dazu tendieren, ihre Informationen mit sehr häufig erscheinenden, breit definierten Begriffen auszudrücken oder mit spezifischeren, weniger häufig auftretenden Begriffen, die aber gerade für die Identifizierung von relevanten Inhalten immens wichtig sind.[130] Die inverse Dokumenthäufigkeit findet man in verschiedenen Varianten. Hier ist die einfachste Form:

$$w_{(i,d)} = \frac{\text{Häufigkeit, mit der i im Dokument d auftritt}}{\text{Anzahl der Dokumente, in denen i vorkommt}}$$

Das Gewicht eines Terms ist dann besonders hoch, wenn es nur wenige Dokumente gibt, in denen er auftaucht, und wenn er gleichzeitig im fraglichen Dokument häufiger vorkommt. Ein Beispiel dazu: Wir betrachten ein Dokument d1 aus einer Kollektion von 1'000 Dokumenten. Die Anzahl aller Dokumente, in denen ein Term i1 gefunden wird, beträgt 650; die Anzahl für einen Term i2

[130] Belew 2000:1

beträgt 80 Dokumente. i1 kommt in d1 50-mal vor und i2 in d1 35-mal. Für die beiden Terme ergeben sich nun die inversen Dokumenthäufigkeiten:

$$idf_{(i1,d1)} = \frac{50}{650} = 0{,}077$$

$$idf_{(i2,d1)} = \frac{35}{80} = 0{,}438$$

Der Term i1 besitzt eine inverse Dokumenthäufigkeit von 0,077; der Term i2 ein Gewicht von 0,438. i2 hat somit eine wesentlich höhere Relevanz im Dokument d1 als der Term i1.

Eine etwas kompliziertere Form der inversen Dokumenthäufigkeit (tf x idf_{id}) kombiniert die Wortfrequenz mit der Dokumentfrequenz und weist einem Wort in einem Dokument im Verhältnis zur Auftretenshäufigkeit des Wortes in dem Dokument und im inversen Verhältnis zur Anzahl der Dokumente in der Kollektion, in denen das Wort mindestens einmal auftaucht, das Gewicht zu.

$$w_{id} = tf_{id} * \log\left(\frac{N}{n_i}\right)$$

wobei

w_{id} = Gewicht der inversen Dokumenthäufigkeit vom Term i im Dokument d
tf_{id} = Frequenz des Terms i im Dokument d
N = Gesamtanzahl aller Dokumente in einer Kollektion C
n_i = Anzahl der Dokumente in C, in denen der Term i enthalten ist

In unserem Beispiel von oben ergeben sich daraus die folgenden Gewichte der inversen Dokumenthäufigkeit:

$$w_{i1d1} = tf_{i1d1} * \log\left(\frac{N}{n_{i1}}\right) = 50 * \log\left(\frac{1000}{650}\right) = 9{,}35$$

$$w_{i2d1} = tf_{i2d1} * \log\left(\frac{N}{n_{i2}}\right) = 35 * \log\left(\frac{1000}{80}\right) = 38{,}39$$

Es ergibt sich wiederum ein wesentlich höheres Gewicht für den Term i2. Das Dokument d1 würde bei einem Retrieval nach dem Suchterm i2 entsprechend relevanter sein als bei einer Suche nach i1. Das Gewicht eines Dokuments in Bezug auf eine Frage, bestehend aus den Indextermen i_1, i_2, ... i_n, errechnet sich dann als Summe der Gewichte w_{in} von den jeweils im Dokument gefundenen Termen i_n. Es werden dadurch Dokumente, die sozusagen ähnlich sind, hoch bewertet. Im Gegensatz zum Bool'schen Ansatz besteht die Retrievalantwort hier nicht in einer ungeordneten, scharf abgegrenzten Menge von Dokumenten, sondern in einer Sortierung der Dokumente nach fallender Übereinstimmung, genannt **Ranking** (*ranking*). Selbst die einfachsten Strategien der Termgewichtung erzielen deutliche Verbesserungen gegenüber einem reinen Bool'schen Retrieval.[131] Mit einer geeigneten Festlegung von einem oberen (maximalen) und einem unteren (minimalen) **Grenz-** oder **Schwellwert** (*threshold value*) für die inverse Dokumenthäufigkeit können zusätzlich Terme aussortiert werden, die im Verhältnis zu allen Indextermkandidaten besonders hohe oder niedrige Frequenzen aufweisen. Auch diese Idee geht auf Luhn zurück.[132]

Eine weitere Gewichtungsformel, die **tfc-Gewichtung** (*term frequency c weighting*) berücksichtigt im Gegensatz zur inversen Dokumenthäufigkeit zusätzlich auch die Länge eines Dokuments. Die Länge jedes Dokuments in einer Kollektion wird normalisiert, sodass die Termfrequenzen in langen Dokumenten nicht stärker bewertet werden als diejenigen in den kürzeren Dokumenten. Die **Ict-Gewichtung** (*Itc-weighting*) verwendet anstelle der blossen Wortfrequenz den Logarithmus der Wortfrequenz, um den Effekt von grossen Frequenzdifferenzen zu reduzieren. Ein weiterer Ansatz, die **Entropie-Gewichtung** (*entropy weighting*), basiert auf informationstheoretischen Ideen und ist die am weitesten entwickelte Termgewichtungsmethode. Die drei Formeln werden hier nicht näher erläutert, denn für die Auswertung der Testdaten in Kapitel 5 wird eine abgewandelte Form der inversen Dokumenthäufigkeit Anwendung finden.

Termgewichtung ist wichtig für die Selektion von guten Indextermen oder für eine bessere Diskriminierung der Indexterme beim Abgleichen mit einer Suchfrage in einer Retrievalumgebung. Die Gewichtung verbessert die Präzision des Retrievals.[133] Sie ist effizient und wird für das Indexieren von grossen und heterogenen Textkollektionen benutzt. Dennoch könnte das Verfahren verfeinert werden. Denn, wie wir bereits gesehen haben, sind nicht alle in einem Text vorkommenden Wörter gute Indexterme, und ihre exakte Auswahl ist von zentraler Bedeutung. Einzelne Wörter sind oftmals zu allgemein, um eine korrekte Inhaltserschliessung zu ergeben. In diesen Fällen sind Phrasen die besseren Indikatoren. Andererseits sind Indexterme manchmal auch zu spezifisch für eine Textrepräsentation, z.B. morphologische Varianten eines Wortes. Die in Texten auftretenden Phänomene Synonymie, Homonymie und Polysemie können Probleme verursachen. In der Analyse und Indexierung der Textausschnitte in Abschnitt 5.3 wird sich

[131] Knorz 1994:15f.; van Rijsbergen 1997:94f.
[132] Kaiser 1996:7
[133] Moens 2000:89

herauskristallisieren, ob eine Variante des Verfahrens der Termgewichtung auch für das automatische Erstellen von Schlagwortregistern für Bücher geeignet ist.

2.5.2.1.3.2 Vektorraummodelle

Häufig werden für die Indexierung einer Information Retrieval-Umgebung auch **Vektorraummodelle** (*vector space models*) eingesetzt. Als Vektorraummodelle werden sie bezeichnet, weil sie vom Ähnlichkeitsbegriff ausgehen und sowohl Suchfrage als auch Dokumente als Vektoren in einem vieldimensionalen Raum repräsentiert werden. Dabei ist jedem Indexterm eine Koordinate zugeordnet. Ähnlichkeit bedeutet dann räumliche Nähe der Vektoren.[134] Dieses heuristische Modell ist ein mathematisch einfaches und gut handhabbares Modell, das viele Ansatzpunkte für differenzierte Ausgestaltungen bietet. Es wird aufgrund der Verwendung der Termgewichtung mit der inversen Seitenhäufigkeit im Teil 2 dieser Arbeit der Vollständigkeit halber hier erwähnt, jedoch nicht näher erläutert.

2.5.2.1.3.3 Probabilistische Retrievalmodelle

Ein weiterer Ansatz der Retrievalforschung ist wahrscheinlichkeitsorientiert: die **probabilistischen Retrievalmodelle** (*probabilistic retrieval*).[135] Die Relevanz eines Dokumentes wird als Zufallsgrösse betrachtet, denn ein Dokument kann für unterschiedliche Benutzer von verschiedenem Nutzen sein (vergleiche Abschnitt 4.4.3). Die Bestimmung der Relevanz steht somit im Mittelpunkt und ist ein Ereignis, das es auf der Basis der verfügbaren Information über das Vorkommen und die Verteilung der Indexterme abzuschätzen gilt. Das Modell hat sich insbesondere beim sogenannten **Relevanz-Feedback** (*relevance feedback*) bewährt, bei welchem versucht wird, die Effektivität eines Information Retrieval-Systems durch die aktive Einbeziehung des Benutzers zu verbessern. Dem Benutzer kommt dabei die Aufgabe zu, die vom System wiedergewonnenen Dokumente nach ihrer Relevanz zu beurteilen. Dieses Urteil wird dann an das System zurückgegeben und darauf aufbauend ein modifiziertes Retrieval durchgeführt.[136] Relevanz-Feedback ist eine Technik der *query expansion*, die der Erweiterung einer Suchanfrage dient. Eine andere Frageerweiterungstechnik basiert auf der Anwendung eines Thesaurus, wodurch Synonyme und andere verwandte Wörter oder Phrasen zur Anfrage hinzugefügt werden. Beide Ansätze erweisen sich als sehr vorteilhaft in Bezug auf die Performanz von Retrievalsystemen.[137]

---

[134] Aas/Eikvil 1999:16-18; Knorz 1994:16f.; Verdejo et al. 1999:7f.
[135] Kaiser 1996:12-19; Knorz 1997:132; van Rijsbergen 1997:93-114
[136] Kaiser 1996:21-37; Knorz 1994:18; Verdejo et al. 1999:9
[137] Lahtinen 2000:95

Wie schon für das Vektorraummodell so gilt auch hier, dass auf das probabilistische Verfahren nicht näher eingegangen wird, es jedoch kurz beschrieben wurde, um die Liste der automatischen Indexierungsverfahren möglichst umfassend zu halten.

2.5.2.1.4 Andere Indexierungsverfahren

Es gibt neben den beschriebenen statistischen Methoden auch noch andere Verfahren zur automatischen Indexierung. Diese werden hier aufgelistet. Für mehr Informationen können Moens[138], Kuhlen[139] und Fugmann[140] konsultiert werden.

- Die Probabilitätsfunktion *Multiple Poisson (nP) Model of Wort Distribution* ist ein statistisches Modell der Wortverteilung in grossen Kollektionen von Volltextdokumenten.
- Ein anderes Modell versucht die Rolle von Diskursstrukturen (*discourse structures*) in den Indexierungsprozess mit einzubeziehen. Es basiert auf dem linguistischen Wissen über Diskursstrukturen, das zur Selektion von inhaltstragenden Termen in einem Text verwendet wird.
- Mit dem Einsatz von Neuronalen Netzen (*neural networks*) verspricht man sich einen Mechanismus, der auf der Grundlage vorhandener Texte selbstständig lernt, wie indexiert werden soll.
- Ein semantisches Netz, bestehend aus Knoten, die für Begriffe oder Begriffsausprägungen stehen, und Kanten zwischen den Knoten, die semantische Beziehungen darstellen, wird als Repräsentationsformalismus für die thematische Dokumentbeschreibung getestet.
- Mit Frame-Sprachen (*frames*) wird versucht, alles Wissen, welches zur Lösung eines Problems benötigt wird, zusammenhängend zu repräsentieren. In einem Musterabgleich (*matching*) zwischen der Frame-Struktur und der Struktur einer Problembeschreibung wird die Relevanz des in einem Frame dargestellten Wissens erkannt.
- Terminologische Logiken bilden eine weitere Art eines Repräsentationsformalismus für Dokumentdarstellungen. Sie legen Begriffsdefinitionen, also Terminologien an.

2.5.2.1.5 Hybrid-Indexierung (automatisch)

Alle bisherigen statistischen Verfahren sind im Grossen und Ganzen ohne Schwierigkeiten zu realisieren.[141] Um den Anforderungen eines Information Retrieval-Systems zu genügen, werden allen kollektionsorientierten Modellen jedoch die oben angesprochenen linguistischen Zusatzschritte, nämlich der Vergleich mit einer Stoppwortliste und die Reduktion der Terme aller

[138] Moens 2000:98-101
[139] Kuhlen 1997
[140] Fugmann 1999
[141] Kaiser 1996:18

Dokumente auf ihre Grund- oder Basisform, vorangestellt. Die Kombination von linguistischen und statistischen Methoden ist heute üblich. Mit dem Einsatz der linguistischen Verfahren bietet sich die Möglichkeit, die eigentlichen Dokumentinhalte besser zu erfassen. Als Vorteil daraus ergibt sich ein höherer Recall (siehe Abschnitt 4.4.4) und eine grössere Benutzerfreundlichkeit. Rein linguistische Verfahren konnten sich bislang nicht durchsetzen.[142] Die Gründe dafür scheinen der umfangreiche Analyseaufwand sowie das Verhältnis von immens grossem Rechenaufwand und unbefriedigendem Ergebnis zu sein.

Das Vorgehen einer automatischen Hybrid-Indexierung läuft etwa nach den folgenden Schritten ab:

1. Identifizierung der Tokens

2. Eliminierung von Funktionswörtern mit Stoppwortlisten

3. Stemmatisierung der verbleibenden Indextermkandidaten (manchmal auch nur Trunkierung[143])

4. Berechnung der Termgewichte anhand eines statistischen Modells

In einer Retrievaloperation erfahren die Wörter der Suchfrage gleichermassen die Schritte 1 bis 3. Die dadurch produzierten Suchterme werden mit den gewichteten Indextermen abgeglichen, und als Suchergebnis werden diejenigen Dokumente an den Benutzer zurückgegeben, deren Indexterme das beste Matching erzielen. Ebenso wird das Ranking der gefundenen Dokumente aufgrund der Gewichtung der Indexterme, die sich mit den Suchtermen decken, ermittelt. Es bestimmt die Reihenfolge, in welcher der Benutzer die gelieferten Dokumente erhält.

2.5.2.2 Zuteilungsmethoden

Wie es schon bei der manuellen Indexierung Extraktions- und Zuteilungsverfahren gab, so unterscheidet man auch bei den automatischen Indexierungsverfahren extrahierende und zuteilende Verfahren. Sie unterliegen denselben Vor- und Nachteilen wie die manuellen Verfahren. In diesem Abschnitt beschäftigen wir uns mit den Zuteilungsmethoden oder der begriffsorientierten Indexierung, also dem automatischen Indexieren unter Einbezug eines kontrollierten Vokabulars.

Die Repräsentation eines Textinhalts mithilfe von natürlichsprachlichen Indextermen aus dem Originaltext bereitet Probleme, allen voran die semantische Ambiguität der Wörter und die Schwierigkeit, diese für Fragestellungen nach allgemeinen Konzepten zu verwenden. Der Gebrauch

[142] Kaiser 1996:20
[143] Trunkierung heisst, dass durch die Beibehaltung oder das Abschneiden einer fixen Anzahl von Zeichen Wortformen systematisch reduziert werden.

eines kontrollierten Indexierungsvokabulars kann gerade diese Probleme bis zu einem gewissen Grad lösen. Deswegen wird auch beim automatischen Indexieren der Schritt unternommen, Indexterme aus einer kontrollierten Indexsprache zu benutzen. Zum Einsatz kommen abermals Thesauri, Klassifikationen und Schlagwortlisten. Auch die maschinelle Erstellung dieser Vokabulare ist von grossem Interesse, da nicht nur die Implementationskosten reduziert werden könnten, sondern auch die Aktualisierung und Pflege sowie die Erweiterung eines Fachbereichs ermöglicht würden.

2.5.2.2.1 Thesaurusbasierte Verfahren

Ein Thesaurus zum automatischen Indexieren hat die Form eines maschinenlesbaren Wörterbuches.[144] Die Originalwörter eines Textes werden mit dessen Hilfe in uniformere Bezeichnungen und allgemeine Konzepte übersetzt. Der Thesaurus stellt auch eine Gruppierung oder Klassifizierung der Terme sowie semantische Verbindungen zwischen den Termen zur Verfügung. Für die maschinelle Desambiguierung von Wortbedeutungen mit einem elektronischen Thesaurus steht für jede Bedeutung eines Wortes eine kurze textuelle Beschreibung zur Verfügung. Auf weitere Eigenschaften sowie die Vor- und Nachteile einer Thesaurusverwendung gehe ich hier nicht noch einmal ein. Es sei dazu auf Abschnitt 2.4.2.2.1 verwiesen.

Thesaurusdeskriptoren sind für die automatische Indexierung von Texten ein effektives Hilfsmittel. Sie können die natürlichsprachlichen Indexterme, die aus einem Text extrahiert wurden, ersetzen. Oder sie können die extrahierten Indexterme bei einem kombinierten Verfahren ergänzen. Dies steht in Analogie zum Gebrauch eines Thesaurus, um in einem Retrievalsystem die Terme einer Frageformulierung mit verwandten Termen zu verbinden.[145] Thesaurusklassen verbessern die Vollständigkeit eines Retrievals. Insbesondere in begrenzten Fachgebieten, wo die darin beschäftigten Personen die Wortbedeutungen teilen, sind Deskriptoren sehr nützlich. Aber für heterogene Textkollektionen müsste man mehr über die gewünschte Form und den Inhalt eines Thesaurus wissen und auch über die zu automatisierenden Prozesse der Wortbedeutungsdesambiguierung.[146]

2.5.2.2.2 Klassifizierende Verfahren

Werden einem Dokument Klassifikationsnotationen oder sonstige Klassenbezeichnungen zugeordnet, so handelt es sich um ein klassifizierendes Indexierungsverfahren. Knorz nennt das

[144] Moens 2000:106
[145] van Rijsbergen 1997:31ff.
[146] Moens 2000:108

Verfahren auch eine **additive Indexierung**.[147] Da einfache Schlagwortlisten gerade im Umfeld der computerbasierten Verarbeitung die maschinellen Möglichkeiten nicht optimal ausnutzen, gleichen Schlagwortlisten zur automatischen Indexierung den Klassifikationen und sind eher so etwas wie Schlagwortkategorisierungscodes (*subject classification codes)*. Die beiden Hilfsmittel werden deshalb in diesem Abschnitt zusammengefasst behandelt.

Die automatische Zuweisung von Schlagwörtern, Notationen oder anderen vordefinierten Kategorien bezeichnet man auch als **Textkategorisierung** (*text categorization*).[148] Systeme, die automatisch Muster in Kategorien aufteilen, nennt man *pattern classifiers* oder kurz *classifiers*. Ein *term text classifier* wird für ein System verwendet, das Schlagwort- oder Klassifikationscodes zuweist.[149] Im Übrigen wird auch die manuelle Zuweisung von Notationen oder Schlagwörtern gelegentlich Textkategorisierung genannt.

Automatische Verfahren im Bereich der Textkategorisierung arbeiten üblicherweise ganz anders als Menschen, indem sie Klassifikationssysteme als Ergebnis einer Datenanalyse (oder Clusteranalyse) selbst erzeugen. Sie fassen solche Dokumente, die sich nur wenig im Wortschatz unterscheiden, zu Klassen zusammen und erzeugen dadurch, dass sie auf höherem Niveau auch Klassen zusammenfassen, eine Klassenhierarchie. Oft wird jede Klasse durch ein prototypisches Dokument repräsentiert. Ein neues Dokument wird dann mit allen Klassenrepräsentanten verglichen und der ähnlichsten Klasse zugeordnet. Die Zuweisung von Schlagwörtern oder Notationen erfordert die Kenntnis der Beziehungen zwischen den Konzepten eines Textes und den Textmerkmalen.[150]

Es wurden diverse Algorithmen zur automatischen Textkategorisierung entwickelt. Hier eine Auswahl der wichtigsten:

- Der Rocchio Algorithmus basiert auf einem Vektorraummodell und ist die klassische Methode für Dokumentrouting oder -filterung.
- Der *Naive Bayes Classifier* wurde konstruiert, indem Trainingsdaten für die Schätzung der Wahrscheinlichkeit jeder Klasse anhand der Merkmale eines neuen Dokuments verwendet wurden. Die Methode ist überraschend effektiv.
- Das Verfahren des *K-nearest Neighbour* (kNN) gliedert die Nachbardokumente eines Dokuments mit den Vektoren der Trainingsdokumente und benutzt die Klassenbezeichnungen der ähnlichsten Nachbarn, um die Klasse eines neuen Dokuments zu bestimmen.
- Ein anderer Ansatz verwendet Entscheidungsbäume (*decision trees*) zur Bestimmung der Relevanz eines Dokuments. Dazu wird ein Dokumentvektor mit dem Entscheidungsbaum, der mithilfe von Trainingsmustern gebildet wird, abgeglichen.

[147] Knorz 1997:134
[148] Aas/Eikvil 1999:12; Moens 2000:103
[149] Moens 2000:111
[150] Knorz 1997:134; Moens 2000:131

- *Support Vector Machines* (SVMs) werden für die Lösung einer breiten Aufgabenpalette der Klassifizierung benutzt. Sie haben sich auch für die Textkategorisierung als nützlich erwiesen.
- Eine weitere Technik kombiniert die Vorhersagen von multiplen Klassifikatoren, um einen einzelnen Klassifikator zu produzieren. Der Prozess wird oft *voting* genannt.

Detaillierte Angaben zur automatischen Textkategorisierung finden sich bei Moens[151], van Rijsbergen[152] und Aas/Eikvil.[153] Das Thema wird hier nicht weiter verfolgt, da es auch im zweiten Teil der Ausführungen nicht zur Anwendung kommt.

2.5.2.3 Probleme der automatischen Indexierung

Die automatische Indexierung kann zufriedenstellende Ergebnisse liefern, wenn die Wortwahl des Suchenden weit gehend mit der Wortwahl der Autoren übereinstimmt, wie es bei den Individualbegriffen meistens der Fall ist, z.B. bei den Namen von Personen, Institutionen und Ländern. Hier sind die Ausdrucksweisen für die Begriffe gut voraussehbar, und ihre mechanisierte Verarbeitung gelingt.[154]

Aber beim automatischen Indexieren stellen sich auch eine Menge Kritikpunkte ein:

- Indexierungsfehler werden normalerweise nicht erkannt, besonders diejenigen, die durch eine Nichtzuweisung von Indextermen entstanden sind. Sie können nur durch Zufälle oder durch ausgedehnte, kostspielige Untersuchungen entdeckt werden.
- Die Übersetzung in eine Indexsprache hängt stark von der Bedeutung der Wörter und Sätze eines Ausgangstextes ab. Doch die Bedeutungsklärungen werden bei den Extraktionsverfahren grösstenteils unterlassen. Auch ein Miteinbeziehen des Kontextes eines Wortes (*co-occurence*) zu dessen Bedeutungsklärung kann nur lückenhaft berücksichtigt werden, da die benachbarten Wörter und Wortgruppen, die hierfür in Betracht gezogen werden, praktisch unbegrenzt mannigfaltig und selbst wiederum mehrdeutig sein können.
- Das eigentliche Erkennen der Essenz eines Textes findet beim automatischen Indexieren nicht statt, auch wenn die Worthäufigkeit als Massstab für die Wichtigkeit eines Begriffes im Text verwendet wird. Die ganz häufigen und die ganz seltenen Wörter werden dabei abgewertet, und nur die in einem Text mittelmässig häufig vorkommenden Wörter werden als Essenzträger anerkannt, wobei die Grenze durch die Festlegung von Schwellwerten willkürlich gezogen werden kann.[155]

[151] Moens 2000:111-131
[152] van Rijsbergen 1979:25-46
[153] Aas/Eikvil 1999:12-20
[154] Fugmann 1999:128
[155] Fugmann 1999:126

- Der Einsatz von Stoppwortlisten ist nicht unumstritten, da gerade der Kontext eines Wortes zu dessen Bedeutungsklärung beitragen kann. Abhilfe können auch Positivlisten, also Listen, in denen die wichtigen Wörter aufgeführt sind, nicht schaffen. Denn auch hier unterbleibt die Klärung der Bedeutung, und zusätzlich werden sämtliche Neuerungen ausgeschlossen.
- Paraphrasierend formulierte Begriffe entziehen sich der automatischen Indexierung. Diese wären jedoch beim Retrieval von allgemeinen Konzepten wesentlich.
- Begriffsverwandtschaften wie Synonymie oder hierarchische Beziehungen zwischen Wörtern werden nicht erkannt, sofern die Wörter nicht morphologisch ähnlich sind. Im Gegenzug werden Homonyme und Polyseme nicht differenziert.

Es ist in erster Linie die Indeterminiertheit der natürlichen Sprache, die für Fugmann das eigentliche Problem der automatischen Indexierung darstellt, das auch nicht überwunden werden kann. Für den Menschen hingegen bedeutet der Umgang mit indeterminierten Prozessen kein Hindernis.[156] Seiner Auffassung nach kann die automatische Indexierung sogar nur bei der Erstellung von Konkordanzen befriedigende Ergebnisse liefern. Dem wäre entgegenzusetzen, dass bei der heutigen Informationsflut eine Indexierung auf rein manueller Basis gar nicht mehr möglich ist und dass dem Nichtvorhandensein einer Indexierung doch immerhin eine automatische vorzuziehen ist. Im Übrigen liefern die gängigen statistischen und kollektionsorientierten Verfahren erstaunlich gute Ergebnisse, und auch manuell ausgeführte Indexierungen sind kein Garant für qualitativ erstklassige Indexe, wie dies die Schlagwortregister gewisser Bücher beweisen. Es scheinen sowohl die manuellen wie auch die maschinellen Techniken Plattformen für Verbesserungen zu bieten.

Damit wäre der Katalog der manuellen und automatischen Indexierungsverfahren abgeschlossen. Es folgt im nächsten Abschnitt ein Vergleich der Verfahren.

2.5.3 Die Indexierungsverfahren im Vergleich

In der Arbeit von Fugmann findet sich eine Gegenüberstellung der Indexierungsverfahren, welche in Kurzform die unterschiedlichen Stärken und Schwächen jedes Verfahrens aufzeigen soll. Ein „–“ bedeutet eine Schwäche, ein „+“ eine Stärke des betreffenden Verfahrens bei einer der Leistungen 1 bis 10. Steht eines dieser Zeichen in Klammern, so besteht die Stärke oder Schwäche des betreffenden Verfahrens nur mit Einschränkung. Das Klassifizieren fällt unter das verbindliche Indexieren.

[156] Fugmann 1999:125

	Indexierungsverfahren					
Leistungen	extrahierend	kontrolliert	verbindlich	frei	manuell hybrid	automatisch
Erkennung und Verdeutlichung der Wortbedeutungen	–	+	+	(+)	+	–
Voraussehbarkeit der Essenzerkennung	–	+	+	–	+	–
Voraussehbarkeit der Essenzdarstellung	–	(+)	+	–	+	–
Auffüllung von Ellipsen	–	+	+	(+)	+	–
Paraphrasen-Lexikalisierung	–	+	+	(+)	+	–
Erfassung von Grafik	–	+	+	+	+	–
Spezifität	(+)	(–)	(–)	(+)	+	(+)
Aktualität	(+)	–	–	(+)	+	(+)
Möglichkeit zur recherchentauglichen Syntax	–	+	+	–	+	–
Wirtschaftlichkeit	strittig					

Tabelle 1: Vergleich der Indexierungsverfahren nach Fugmann[157]

Auffallend an der Aufstellung ist, dass die Wirtschaftlichkeit bei allen Verfahren strittig ist. Fugmann begründet dies dadurch, dass in Anbetracht der geringen Qualität, die mit billiger Indexierung (insbesondere mit extrahierendem und freiem Indexieren) erreichbar ist, solche Verfahren ebenso gut auch unwirtschaftlich sind, während die aufwändigen und kostspieligen Verfahren auf längere Sicht hin auch als wirtschaftlich gelten können.[158]

Interessant ist auch, dass auf das Indexieren von Grafiken Wert gelegt wird. Sie entzieht sich anhand dieser Gegenüberstellung dem extraktionsbasierten und automatischen Indexieren jedoch gänzlich. Bei der Identifikation von nicht-interpretationsbedürftigen Grafiken (Landkarten, Sternbilder, Fingerabdrücke usw.) hingegen kann die Automatik laut Fugmann durchaus nützliche Dienste leisten.[159]

[157] Fugmann 1999:131
[158] Fugmann 1999:132
[159] Fugmann 1999:133

Beim Extraktionsverfahren sind als einzig Positives die Spezifität und die Aktualität zu vermerken, und dies auch nur mit Einschränkung. Die Zuteilungsverfahren *kontrolliert* und *verbindlich* leiden dagegen unter mangelnder Spezifität und Aktualität ihres Wortschatzes. Dafür aber besteht bei ihnen die Möglichkeit einer recherchentauglichen Syntax, durch welche die Spezifität verbessert werden kann, die bei beiden Verfahren eingeschränkt worden ist. Beim verbindlichen Indexieren ist die Voraussehbarkeit der Essenzdarstellung besser als beim lediglich kontrollierten Indexieren. Ausserdem schneidet beim verbindlichen Indexieren die Spezifität auch besser ab als beim kontrollierten, weil man sich offenbar darauf verlassen kann, dass die Indexierer von der Spezifität des Wortschatzes beim verbindlichen Indexieren tatsächlich Gebrauch machen.

Beim freien Indexieren werden fast alle Positiva mit Einschränkung gesehen. Immerhin kann gutes freies Indexieren die Wortschatzschwäche vom kontrollierten und verbindlichen Indexieren ausgleichen. Vor allem ist freies Indexieren bei Individualbegriffen recht effektiv. Das manuelle Hybrid-Indexieren, d.h. die Kombination von verbindlichem und freiem Indexieren, wird von Fugmann ausschliesslich positiv beurteilt, während die automatische Indexierung in all ihren Varianten im Vergleich dazu überwiegend negativ bewertet wird. Die unüberwindliche Hürde ist für Fugmann die Indeterminiertheit der natürlichsprachlichen Ausdrucksweise und die daraus folgende Nichtprogrammierbarkeit einer befriedigenden algorithmischen Verarbeitung.[160] Positiv bewertet bei der automatischen Indexierung sind hingegen Spezifität und Aktualität, wenn auch mit Einschränkung.

Es ist offensichtlich, dass Fugmann der extrahierenden und automatischen Indexierung gegenüber ziemlich negativ eingestellt ist. Wobei zu erwähnen ist, dass das übliche automatische Indexieren mit den linguistischen Methoden und der Termgewichtung in der Tabelle stellvertretend für die gesamte Palette der automatischen Indexierungsverfahren steht. Die Volltextspeicherung findet gar keinen Platz. Immerhin gesteht Fugmann ein, dass computerunterstütztes Indexieren zu Registern von hoher Qualität führt.[161] Man müsste bei einem Vergleich der Indexierungsverfahren wohl auch eine Differenzierung innerhalb der automatischen Methoden vornehmen. Und meiner Auffassung nach ist es unerlässlich, bei einer Beurteilung auch den Zweck der Indexierung mit einzubeziehen. Denn je nachdem, welche Aufgabe in welchem Bereich eine Indexierung zu erfüllen hat, also ob sich das spätere Retrieval beispielsweise in einem Buch oder im Internet abspielt, kann ein und dasselbe Verfahren äusserst nutzbringend sein oder aber überhaupt nicht. Ausserdem sollte der Leistung der Wirtschaftlichkeit ein grösserer Stellenwert zukommen.

Die von Fugmann bereitgestellte Gegenüberstellung steht in krassem Gegensatz zu den Ausführungen anderer Autoren im Umfeld des Indexierens. Van Rijsbergen hält beispielsweise fest, dass eine Indexierung mit einem unkontrollierten Vokabular oder eine Extraktionsindexierung Retrievalergebnisse erreicht, die mit den Ergebnissen der Indexierungsverfahren mit einem streng

[160] Fugmann 1999:133
[161] Fugmann 1999:174

kontrollierten Vokabular vergleichbar sind.[162] Knorz berichtet von einer automatischen thesaurusbasierten Indexierung, die in einem Retrievaltest Ergebnisse erzielen konnte, die nur graduell ungünstiger, in Teilbereichen sogar günstiger waren als die Resultate der manuellen Indexierung.[163] Eine andere Reihe von Tests zum Vergleich von manueller thesaurusbasierter Indexierung und automatischer Indexierung (Volltextspeicherung und Termgewichtung) legt für Knorz den Schluss nahe, dass die automatische Indexierung nicht nur ebenbürtige, sondern vermutlich sogar die besseren Retrievalergebnisse bringt. Als Gesamturteil konstatiert er, dass die automatische Indexierung brauchbar, eine Kombination von automatischer und manueller Indexierung deutlich besser als jede einzelne von beiden ist.[164] Selbst unter realistischen Testbedingungen hat sich gezeigt, dass statistische Retrievalansätze gute Ergebnisse produzieren, insbesondere beim Einsatz von Relevance-Feedback-Techniken erweisen sich vollautomatische Retrievalverfahren den manuellen Techniken gegenüber als deutlich überlegen. Im Bereich der automatischen Extraktionsverfahren erzielen zusätzliche linguistische Verfahren zur Wortformennormierung und Kompositazerlegung die besseren Resultate als die Volltextspeicherung oder die rein statistischen Modelle.[165]

Der Grund für die weit auseinander laufenden Meinungen könnte in dem extrem breiten Anwendungsbereich des Indexierens liegen. Erneut möchte ich darauf hinweisen, dass deshalb der Zweck einer Indexierung bei deren Beurteilung immer berücksichtigt werden sollte. Ferner ist zu beachten, dass in der heutigen Zeit der Informationsmassen und der Konfrontation mit einer riesigen Menge produzierter Dokumente eine Informationserschliessung auf manueller Basis kaum mehr durchführbar ist.

[162] van Rijsbergen 1997:13
[163] Knorz 1997:139
[164] Knorz 1994:28
[165] Kaiser 1996:20; Knorz 1997:139

3 Das Schlagwortregister

In diesem Kapitel geht es um Schlagwortregister, Schlagwörter und um die Erstellung von Schlagwortregistern für Bücher. Im Prinzip gelten alle bisherigen Ausführungen über das Indexieren von Dokumenten und Texten auch für Schlagwortregister. Dennoch sollen die Besonderheiten der Schlagwortregister hervorgehoben werden. Zur Einleitung dienen die folgenden Zitate:

> „There is no such thing as a good book if it has a poor index or no index at all. Such books are incomplete and are similar to those books published with errors, like blank pages where text should be, or with an upside-down page."[166]

> „A book without an index is incomplete."[167]

3.1 Das Schlagwortregister und seine Funktionen

Wie schon einmal erwähnt, so verstehen die meisten Leute unter einem Index einen **Buchindex** (*book index, back-of-the-book-index*). Sie machen einen grossen Teil der Indexproduktion aus und sind im Wesentlichen Listen von Wörtern, die aufgrund spezifizierter Kriterien meist alphabetisch geordnet sind und ganz hinten in einem Buch stehen. Für jeden Eintrag gibt die Indexreferenz eine Seitenzahl, manchmal auch eine Paragrafennummer, des Gegenstandes oder eine mit dem Gegenstand verknüpfte Bezeichnung an. Ein Buchindex (im Folgenden auch synonym für *Schlagwortregister*, *Schlagwortverzeichnis* oder *Register* verwendet[168]) ist laut Mulvany eine strukturierte Aufeinanderfolge von synthetischen **Zugangspunkten** oder **-stellen** (*access points*) zu allen im Text enthaltenen Informationen, die aus einer gründlichen und vollständigen Textanalyse resultieren.[169] Cleveland/Cleveland beschreiben ein Schlagwortverzeichnis als ein in sich abgeschlossenes Retrievalsystem, bei welchem die Informationsdatenbank zwischen den Deckeln des Buches liegt.[170] Bei Hahn sind Buchindexe „inhaltliche Verdichtungen zur Charakterisierung eines Textes durch Schlagwörter".[171]

Ein Schlagwortverzeichnis fungiert als Zeiger zu den in einem Buch enthaltenen Informationen, sodass ein Leser nicht das ganze Buch lesen oder nochmals lesen muss. Es sortiert und klassifiziert

[166] Bonura 1994:14
[167] Cleveland/Cleveland 1990:142
[168] Bei der Verwendung der Terme *Schlagwortregister*, *Schlagwortverzeichnis*, *Register* oder *Buchindex* kann sowohl ein reines Sachverzeichnis oder eine Mischung aus Sach- und Namenverzeichnis gemeint sein.
[169] Mulvany 2000:4
[170] Cleveland/Cleveland 1990:125
[171] Hahn 1986:196

präzise den Inhalt eines Buches und stellt ihn in einer Form dar, die einem Indexbenutzer den direkten Zugang zu bestimmten gewünschten Einzelheiten erlauben. Die effiziente Lokalisierung von Informationen wird durch die strukturierte Anordnung des Registers ermöglicht. Dennoch ist ein Buchindex kein Inhaltsverzeichnis, obwohl dessen Zweck ebenfalls das Bezeichnen des Inhalts ist. Ein Buchindex folgt nicht der Wortreihenfolge und der konzeptuellen Textgliederung eines Autors. Er fokussiert ein generelles Thema und besitzt einen klar definierten Anfang und ein ebensolches Ende. Ein weiterer Unterschied zwischen einem Buchindex und anderen Indextypen besteht darin, dass jeder Buchindex so einzigartig ist wie das Buch selbst. Er ist nur nützlich für das eine Buch und verändert sich im Allgemeinen nur im Zusammenhang mit einer neuen Auflage. Ein Buchindex ist somit in gewissem Sinne „persönlicher“ als andere Indexformen. Normalerweise werden Buchindexe von Einzelpersonen gemacht im Gegensatz zu anderen Indexen, die von vielen Indexierern über eine relativ lange Zeitspanne geschaffen werden.[172] Es gilt auch für Buchindexe, dass der Index kein Ersatz für die Informationen im Buch ist, sondern lediglich ein Verweis auf die Informationen.

Der Zweck des gedruckten Registers zu einem Buch besteht darin,[173]

- die in einem Buch erörterten Gegenstände klassifiziert und sortiert darzustellen und durch Schlagwörter sowie die Angabe von Seitenzahlen leicht auffindbar zu machen,
- zwischen echter Information und einer lediglich beiläufigen Erwähnung des gesuchten Gegenstandes zu unterscheiden und Letztere auszuschliessen,
- verwandtschaftliche Beziehungen zwischen den Schlagwörtern wiederzugeben,
- Informationen zu einem Thema, die durch die Anordnung des Buchtextes verstreut sind, zu gruppieren,
- den Benutzer beim Nachschlagen von Informationen zu unterstützen und ihn zu den entsprechenden Schlagwörtern des Registers sowie schliesslich zu den die Schlagwörter behandelnden Textpassagen mittels Seitennummerangaben hinzuführen,
- dem Suchenden Hilfe bei der Bestimmung zu bieten, ob ein gesuchtes Thema in einem Buch vorkommt, damit das Buch zurückgewiesen werden kann, wenn das Thema überhaupt nicht zur Sprache kommt.

Es ist eine Tatsache, dass ein Fachbuch, das kein Schlagwortverzeichnis besitzt, weniger gekauft und gelesen wird als eines mit Index. Denn die Leser möchten an spezifische Informationen in einem Buch herankommen, ohne das ganze Buch lesen oder erneut lesen zu müssen, oder sie verwenden ein Buch als Nachschlagewerk.[174] Wenden wir uns nun dem Inhalt eines Registers zu, das für ein Buch so wichtig zu sein scheint.

[172] Cleveland/Cleveland 1990:125
[173] Bonura 1994:14; Fugmann 1999:143; Mulvany 1994:5f.
[174] Cleveland/Cleveland 1990:126

3.2 Die Bestandteile eines Schlagwortregisters

Ein Schlagwortregister besteht im Wesentlichen aus einer Liste von üblicherweise alphabetisch angeordneten **Einträgen**. Daneben existiert in vereinzelten Fällen auch eine **Vorbemerkung (***introductory note*, *headnote***)**, die Auskunft über die bei der Herstellung des Registers befolgten Prinzipien geben. Beispielsweise wird darin angegeben, welche Teile eines Buches nicht indexiert worden sind (Inhaltsverzeichnis, Vorwort, Fussnoten, Bibliografie, Anhang usw.) oder wie die alphabetische Reihenfolge der Schlagwörter zustande kam, insbesondere die Behandlung von Umlauten, Sonderzeichen und Leerzeichen. Eine Vorbemerkung sollte auf keinen Fall fehlen, wenn ein Index vom üblichen Präsentationsformat abweicht oder wenn sich der Index nicht auf alle Bereiche eines Buches bezieht.[175] In den Büchern der Sekundärliteratur zum Thema des Indexierens, die für die Erstellung dieser Arbeit konsultiert wurden, habe ich nur in 2 von 23 Fällen eine Vorbemerkung angetroffen, obschon die Register zum Teil deutliche Unterschiede aufweisen.

3.2.1 Die Einträge

Jeder **Eintrag** (*entry*) in einem Register umfasst:

a) das Schlagwort (im nachfolgenden Beispiel in Fettdruck) mit seinen Fundstellen
b) gegebenenfalls die Untereinträge mit Fundstellen
c) gegebenenfalls die Verweisungen vom Typ „siehe auch“ oder „siehe“

Ein Beispiel:

> **Synonymie**
> *siehe auch* Begriffsverwandtschaften
> Definition 81
> Beispiele 82
> im Thesaurus 174, 328
> verknüpft durch Verweisung 185

Spricht man im Zusammenhang mit Registern von der Anzahl von Einträgen eines bestimmten Verzeichnisses, so ist damit automatisch auch die Anzahl der Schlagwörter dieses Registers gemeint.

Wir setzen uns in den folgenden Abschnitten genauer mit diesen Bestandteilen eines Registereintrags auseinander.

[175] Mulvany 1994:69

3.2.1.1 Das Schlagwort

Das **Schlagwort** (*main heading*) steht auf der obersten Zeile eines Eintrages. Im Englischen existieren diverse Bezeichnungen: *main heading*, *access point*, *entry*, *index entry*, *heading*, *subject heading*. Mulvany schlägt vor, den Term *main heading* zu verwenden, um Verwirrungen vorzubeugen, die besonders bei der Verwendung von *entry* oder *index entry* entstehen können. Das Schlagwort zusammen mit dem ganzen dazugehörigen Block von Informationen nennt sie *entry*. [176] Wir wollen diese Begriffsfestlegungen übernehmen.

Ein Schlagwort kann aus einem einzelnen Wort oder aus mehreren Wörtern bestehen, sogenannte **Mehrwortterme** oder **Multiwortterme** (*multiple terms*, *multi-word terms*).[177] Dasjenige Wort, das in einem Mehrwortterm an der ersten Stelle steht und den Platz des Schlagwortes bei einer alphabetischen Einordnung bestimmt, bezeichnet man als **Ordnungswort**.[178] Wenn ein Leser nach Informationen zu einem Thema sucht, so wird er das Thema auf der Ebene der Schlagwörter, bei mehrgliedrigen Schlagwörtern anhand des Ordnungswortes zu lokalisieren versuchen. Somit sind es die Schlagwörter, die den primären Zugang zu den Informationen im Register ermöglichen. Schlagwörter sind deshalb für das Wiederfinden von Informationen sehr entscheidend und sollten so spezifisch wie möglich sein und einen direkten Bezug zu den Konzepten im Text herstellen. Spezifität kann durch die Verwendung von Mehrworttermen erreicht werden. Wenn ein einzelnes Wort für eine Begriffsbenennung nicht existiert, sollte eine passende Kombination gesucht werden. Ferner bestimmt die Art der Schlagwörter auch die Art der Untereinträge.

Was aber steht auf der obersten Zeile an der Stelle des Schlagwortes? Für Fugmann ist ein Schlagwort auf der Signifié-Seite (Inhaltsseite) eine natürlichsprachliche Ausdrucksweise, welche einen in einem Buch erörterten Begriff prägnant darstellen soll, unabhängig davon, wie der Begriff im Text ausgedrückt worden ist.[179] Auch Ruge betont die Funktion eines Schlagwortes, das für sie ein Begriff ist, der zur Charakterisierung der Themen und Inhalte eines Buches dient.[180] Schlagwörter und Deskriptoren sollten so weit gehend der Fach- und Umgangssprache entsprechen, dass sie dem Suchenden oder Indexierer leicht einfallen. Was die morphologischen Merkmale auf der Signifiant-Seite (Ausdrucksseite) von Schlagwörtern angeht, so ist ein Schlagwort laut Mulvany sehr oft ein Substantiv oder ein von einem Adjektiv begleitetes Substantiv. Alleinstehende Adjektive oder Partikeln eignen sich nicht als Schlagwörter. Nur wenn sie selbst das Thema einer Erörterung bilden, wird eine Ausnahme toleriert.[181] Anstelle der adjektivischen Formen („löslich“) sollten möglichst immer die substantivischen („Löslichkeit“) verwendet werden.[182] Es wird sich

[176] Mulvany 1994:13f.
[177] Lahtinen 2000:168; Schneider 2000:6
[178] Fugmann 1999:163
[179] Fugmann 1999:221
[180] Ruge 1995:240
[181] Mulvany 1994:71, 82
[182] Fugmann 1999:70

herausstellen, ob sich diese Aussagen von Mulvany bewahrheiten, denn wir werden der Frage nach den Wortkategorien von Schlagwörtern im zweiten Teil der Arbeit durch eine detaillierte Auseinandersetzung mit den realen Schlagwortverzeichnissen nachgehen (vergleiche Abschnitt 5.2.2).

Oftmals sind bei einem Schlagwort viele Textpassagen zu nennen, in denen das Thema des Schlagwortes erörtert ist, und zwar in unterschiedlichen Zusammenhängen. Hat ein Suchender aber nur einen ganz bestimmten Zusammenhang im Auge, ist es zeitraubend und lästig, wenn man immer nur durch Aufschlagen der betreffenden Seite erkennen kann, ob das Schlagwort der Fragestellung dort in dem gewünschten Zusammenhang steht. Es ist dann hilfreich, wenn die Fundstellen unter einem Schlagwort durch Untereinträge interpretiert und aufgegliedert sind.

3.2.1.2 Die Untereinträge

Die **Untereinträge** (*subheadings*, *subentries*) sind diejenigen Zeilen in einem Register, die dem Schlagwort unmittelbar folgen und die zusammen mit dem Schlagwort einen Eintrag bilden. Sie dienen der Gliederung eines Schlagwortes in verschiedene Unterthemen oder Aspekte und stehen mit ihm in einer logischen Relation. Im Beispiel sind die fünf eingerückten Zeilen unter dem Schlagwort „Synonymie“ die Untereinträge:

Synonymie
siehe auch Begriffsverwandtschaften
Definition 81
Beispiele 82
im Thesaurus 174, 328
verknüpft durch Verweisung 185

Den weit verbreiteten Ausdruck *Unterschlagwort* möchte ich vermeiden, weil es sich bei den Untereinträgen nicht um eine Art von Schlagwort handelt, sondern um frei formulierte Erläuterungen, die sehr vielfältig sein können, was die Voraussehbarkeit des Wortlautes der Untereinträge verhindert. Mithilfe von Untereinträgen kann man aber in visuell wirksamer Weise zum Ausdruck bringen, in welchem Zusammenhang das betreffende Schlagwort im Text eingebettet ist. Es gilt die Regel, dass Untereinträge erst ab vier bis fünf Fundstellen bei einem Schlagwort gemacht zu werden brauchen.[183] Ihre Verwendung sollte dennoch im Hinblick auf die Platzknappheit in einem gedruckten Buch nicht übertrieben werden.

Untereinträge sind für den Suchenden sehr zeitsparend. Sie beschreiben in Kurzform den Zusammenhang, in welchem ein Schlagwort im Buchtext anzutreffen ist. Sie sollen wie die

[183] Mulvany 1994:77

Schlagwörter selbst informativ und prägnant sein. Die Darstellung der Begriffsverknüpfung in Gestalt von Untereinträgen ist für den Leser hochgradig informativ. Alle Arten von Termbeziehungen, die mit einer Indexsprachengrammatik erfasst werden können, lassen sich durch geschickt formulierte Untereinträge für einen Benutzer leicht verständlich ausdrücken.

3.2.1.3 Die Fundstellen

Die **Fundstellen** oder **Referenzstellen** (*reference locators*) sind üblicherweise Seitenzahlangaben. Sie geben einem Benutzer die Seiten an, wo das Thema seiner Fragestellung behandelt wird. Sie sollen möglichst präzise sein und auf diejenigen Passagen hinweisen, die für ein Schlagwort einschlägig sind. Eine allgemeine Regel verlangt, dass pro Schlagwort oder Untereintrag nur fünf Fundstellen angebracht sind, ansonsten muss eine weitere Differenzierung des Eintrags gemacht werden.[184]

Anstelle der Seiten können in einem Buch auch die Absätze nummeriert werden, und es sind diese Absatznummern, auf welche im Register dann verwiesen wird (in äusserst seltenen Fällen werden auch Zeilennummern angegeben). Das Verfahren setzt die Mitwirkung der Buchautoren voraus, hat aber den Vorteil, dass ein Register schon vor der Durchführung des Seitenumbruchs erstellt werden kann. Dass bei den späteren Korrekturen eine Verschiebung der Absatznummern eintritt, kann leicht vermieden werden, indem man für ergänzende Absätze Dezimalzahlen verwendet.[185] Diese lassen sich an jeder beliebigen Stelle ermitteln. Beispielsweise wird zwischen den Absatznummern 11 und 12 die Absatznummer 11,5 eingeschoben. Ähnlich unproblematisch ist der eventuelle Wegfall ganzer Absätze.

In manchen Registern geben die Fundstellen **mehrteilige Referenzzahlen** an, sogenannte *multipart page numbers*.[186] Dabei wird beispielsweise die Kapitelnummer in die Fundstelle mit einbezogen. Jedes Kapitel beginnt mit der Seite 1. Eine Referenzstelle „5-13“ müsste somit als Seite 13 des Kapitels 5 gelesen werden. Mehrteilige Referenzstellen werden eher selten verwendet, hauptsächlich in Benutzerhandbüchern für Software.

Man kann in den Fundstellen ein Thema, das sich über mehr als eine Seite erstreckt, zusammenhängend abhandeln („35-38“) oder aber mit Unterbrechungen („35, 36, 37, 38“). Die Feststellung dieses Unterschieds ist jedoch ausgesprochen subjektiv und für den Suchenden meist uninteressant. Er wird in beiden Fällen die Seiten 35 bis 38 lesen. Sehr häufig wird nur der Beginn der Erörterung eines Schlagwortes und die Fortsetzung derselben mit „f.“ („folio“ in der Bedeutung „auch nachfolgende Seite“) oder mit „ff.“ („auch auf den zwei nachfolgenden Seiten“) angedeutet,

[184] Mulvany 1994:86
[185] Fugmann 1999:147
[186] Mulvany 1994:92

was die effiziente Registerbenutzung etwas einschränkt. Es obliegt dann dem Benutzer, das Ende der für ihn wichtigen Passage zu entdecken. Meistens werden „35f." und „35, 36" in der gleichen Bedeutung von separaten Referenzstellen auf den Seiten 35 und 36 benutzt; „35ff." und „35, 36, 37" meinen separate Referenzstellen auf den Seiten 35, 36 und 37. Zusammenhängende Themendiskussionen über mehrere Seiten hinweg nehmen die Form „35-36" bzw. „35-37".[187]

Spalten- und **Quadrantenangaben** (*column and quadrant identifiers*) erleichtern die Suche nach einer Referenzstelle in grossen Büchern. Bei Spaltenangaben wird nicht nur auf die Seite, sondern auch auf die entsprechende Spalte hingewiesen. Eine imaginäre Einteilung einer Seite ergibt vier Quadranten, die mit Buchstaben identifiziert werden. Der Quadrant *a* ist der obere linke, *b* der untere linke, *c* der obere rechte und *d* der untere rechte Quadrant. Eine Fundstelle „35b" verweist auf eine Textpassage, die sich auf der Seite 35 etwa im unteren linken Bereich befindet. Das System kann nur sinnvoll genutzt werden, wenn es in den Vorbemerkungen zum Register erklärt wird.

3.2.1.4 Die Querverweise

Die **Querverweise** *(cross-references*) vom Typ „siehe auch" (*see also*) und „siehe" (*see*) sind keine Verweise zu Textstellen; sie dienen der Navigation eines Benutzers innerhalb des Schlagwortregisters und verbessern so dessen Nutzungsmöglichkeiten. Ein „siehe"- oder „siehe unter"-Verweis bringt zusammengehörige in einem Index verstreute Informationen zusammen. Er versucht, die Sprache eines Benutzers sozusagen vorauszusehen und verbindet die Sprache des Buchautors mit derjenigen des Suchenden. Synonyme oder ähnliche Begriffe können mit „siehe"-Querverweisen erfasst werden. Als Gegenstück zu der Verweisung „siehe" wird gelegentlich die Verweisung „benutzt für" angebracht, allerdings eher selten.

Die primäre Funktion eines „siehe auch"- oder „siehe auch unter"-Verweises ist die Verknüpfung von hierarchischen und assoziativen Relationen zwischen den Einträgen sowie der Hinweis auf zusätzliche Informationen unter einem anderen Schlagwort. Es ist darauf zu achten, dass die **Rückbezüglichkeit** (*reciprocity*) zwischen den „siehe-auch"-Verweisungen gewährleistet ist.

Gewisse Situationen machen die Anbringung von Verweisen fast unerlässlich, auch wenn keine absoluten Regeln gelten:

- Unterschiede in der Rechtschreibung eines Terms („Photografie *siehe* Fotografie")
- fremde Wörter oder Namen („Brittany *siehe* Bretagne")
- altertümliche Ausdrücke („Eidam *siehe* Schwiegersohn")
- umgangssprachliche Ausdrücke („Knete *siehe* Geld")

[187] Mulvany 1994:91

- Synonyme, Quasi-Synonyme und Antonyme („Sem *siehe* Merkmal, semantisches“; „Tiefenstruktur *siehe auch* Oberflächenstruktur“)
- Hyponymie und Hyperonymie („Insekt *siehe* Biene“; „Biene *siehe* Insekt“)
- Namenvariationen (“Becket, Thomas *siehe* Thomas à Becket, Saint”)
- Abkürzungen und Akronyme („UNO *siehe* Organisation der Vereinten Nationen“)

Es ist zu beachten, dass Querverweise in nur einem einzigen Schritt zu zusätzlicher Information führen sollten. Ein Benutzer, der erst nach zwei oder drei Zwischenschritten zu einem für ihn vielleicht wichtigen Schlagwort gelangt, wird unnötig verärgert. Sogenannte **blinde** sowie **zirkuläre Verweise** (*blind and circular cross-references*), die den Suchenden ins Leere (nach Nirgendwo) oder in ein Hin und Her zwischen den Schlagwörtern schicken (siehe Titelblatt), beeinträchtigen dessen Vertrauen in das Register und in das gesamte Buch.

3.3 Schlagwortregistertypen

Schlagwortregister lassen sich anhand der Art der Schlagwörter in Namenregister oder Sachregister aufteilen. Mit Namenregister sind nicht die im Abschnitt 2.2 vorgestellten Wort- und Namenindexe gemeint. Es ist im Umfeld der Buchindexierung dennoch üblich, von Namenregistern zu sprechen; deshalb möchte auch ich den Begriff so verwenden.

3.3.1 Namen- und Sachregister

Werden in ein Register die Namen der Autoren von der Sekundärliteratur, die in einem Buch verwendet wurden, eingetragen, so handelt es sich um ein **Namenregister** (*name index*). Es gibt zur Bezeichnung derer zahlreiche Begriffe: Namen(s)register, Namen(s)verzeichnis, Namen(s)index, Autorenregister, Autorenverzeichnis, Autorenindex, Personenregister, Personenverzeichnis, Personenindex. Die Namen können auch Organisationen oder Institutionen bezeichnen. Bei der Erstellung von Namenregistern sind Genauigkeit, Konsistenz und Regeleinhaltung von Bedeutung.

Ein **Sachregister** (*subject index*) listet die Schlagwörter, welche die in einem Buch behandelten Gegenstände, Konzepte oder Themen benennen, auf. Auch hier existieren eine Menge von Bezeichnungen: Sachregister, Stichwortregister, Schlagwortregister, Sachverzeichnis, Stichwortverzeichnis, Schlagwortverzeichnis, Sachindex, Stichwortindex, Schlagwortindex. Die Herstellung von Sachregistern ist schwieriger als die von Namenregistern.

Sehr oft werden Namen- und Sachverzeichnisse in einem einzigen Register vereint. Diese Kombinationen werden trotzdem auch Sachverzeichnisse, Sachindexe, Schlagwortregister,

Stichwortverzeichnisse etc. genannt. Es ist in der Fachwelt des Indexierens strittig, ob man Autoren- und Sachregister getrennt lassen sollte oder ob man besser beide in einem einzigen alphabetischen Schlagwortregister zusammenfasst. Für eine Zusammenfassung spricht die Tatsache, dass die Registerbenutzer zuweilen gar nicht bemerken, dass ein Buch zwei getrennte Register hat. Sie suchen dann die Autoren im Sachregister, finden sie dort nicht und beklagen dies als einen Mangel des Registers. Ich lege mich darauf fest, dass ich unter dem Begriff *Sachregister* ein Schlagwortregister meine, das auch Namen enthalten kann.

Gliedert man die Schlagwortverzeichnisse nach einem anderen Gesichtspunkt, nämlich nicht auf der Grundlage der Art der Schlagwörter, sondern aufgrund der Anordnung der Einträge, so unterscheidet man alphabetische und systematische Register.

3.3.2 Das alphabetische Schlagwortregister

Das **alphabetische Schlagwortregister** (*alphabetic subject index*) ist das, was man in den Büchern eigentlich immer vorfindet. Es listet die Schlagwörter in einer alphabetischen Reihenfolge auf und ist für den Benutzer einfach zu verwenden, da jeder das Prinzip von alphabetisch geordneten Verzeichnissen kennt. Hier ein Ausschnitt aus einen alphabetischen Register als Beispiel:

Abwärtsrecherche
siehe auch Recherche allgemein
ermöglicht durch Systematik 208, 209
bei Insekten 104
mit Klassifikationen 105

Autorenrecherche
siehe auch Recherche allgemein
Variante von Known Item Search 48

.
.
.

Computerrecherche
siehe auch Recherche allgemein
Auffindbarkeit von Zeichenfolgen 510
Unabhängigkeit vom Ordnungswort 255

.
.
.

Recherche allgemein
siehe auch Abwärtsrecherche, Autorenrecherche, Computerrecherche, Known Item Search
Mechanismus von 288
lückenhaft bei fehlenden Suchbedingungen 112
behindert durch Präkombinationen 142
im Volltext 89, 100

Das alphabetische Schlagwortregister leidet darunter, dass die Verwandtschaftsbeziehungen zwischen den Schlagwörtern nur in unübersichtlicher Form angeboten werden, nämlich nur durch die Verweisungen auf andere Schlagwörter. Was alles zu einem Thema gehört, kann weit verstreut sein. Wenn die Zahl der Verweisungen allzu gross ist, wird der Fragesteller abgeschreckt und nimmt die Mühe nicht auf sich, alle konsequent zu verfolgen. Zu viele Querverweise können die effiziente Nutzung eines Registers beeinträchtigen. Wird die Darstellung der Zusammenhänge in einem Buchindex als zu bedeutend angesehen, muss auf ein systematisches Register zurückgegriffen werden.

3.3.3 Das systematische Schlagwortregister

Das gleiche Material des alphabetischen Registers von oben hat als **systematisches Register** (*systematic subject index*) die folgende Struktur:

Recherche allgemein
siehe auch Abwärtsrecherche, Autorenrecherche, Computerrecherche, Known Item Search
Mechanismus von 288
lückenhaft bei fehlenden Suchbedingungen 112
behindert durch Präkombinationen 142
im Volltext 89, 100

Abwärtsrecherche
siehe auch Recherche allgemein
ermöglicht durch Systematik 208, 209
bei Insekten 104
mit Klassifikationen 105

Autorenrecherche
siehe auch Recherche allgemein
Variante von Known Item Search 48

Computerrecherche
siehe auch Recherche allgemein
Auffindbarkeit von Zeichenfolgen 510
Unabhängigkeit vom Ordnungswort 255

Die Gestalt des systematischen Registers hat ihr Vorbild in der systematischen Katalogisierung, ist aber bisher kaum in der Indexierungspraxis anzutreffen.[188] Dieses Register ist hilfreich, wenn ein Benutzer Schwierigkeiten hat, für sein Fragethema ein treffendes Schlagwort zu ersinnen. Die Systematik führt dann mit hoher Wahrscheinlichkeit zu dem gesuchten Schlagwort, falls vorhanden. Auch hier ist in der Fachwelt Skepsis anzutreffen, vor allem weil man dem Suchenden nicht zutraut, dass er von einem systematischen Register Gebrauch machen würde. Jedoch lässt sich die Nutzung

[188] Fugmann 1999:146

der Systematik dadurch sehr erleichtern, dass man einen Zugang von einem zusätzlichen alphabetischen Register her eröffnet.

Das wenig gebräuchliche systematische Register ist dem alphabetischen dadurch überlegen, dass es die verwandten Schlagwörter in Nachbarschaft (**Juxtaposition**) bringt und dass es auf diese Weise gegenüber dem alphabetischen Register das zeitraubende Nachschlagen an vielen weit auseinander liegenden Stellen erspart. Viele verwandtschaftlichen Beziehungen zwischen den einzelnen Schlagwörtern kommen allein schon durch ihre Anordnung in Nachbarschaft zueinander zum Ausdruck. Das gilt vor allem für die hierarchisch miteinander verwandten Begriffe. Dies trägt sehr zur Übersichtlichkeit und Platzersparnis bei. Das blosse Auf- und Absteigen eines Suchenden in der Systematik erfordert keine Verweisungen. Trotzdem wirkt ein systematisches Register fremdartig.

Optimal für den Suchenden wäre es, wenn ihm bei einem Nachschlagewerk sowohl ein alphabetisches als auch ein systematisches Schlagwortregister zur Verfügung stünde, und zwar in einer ausführlichen Form mit zahlreichen differenzierenden Untereinträgen. Im Buchdruck stehen einem solchen Angebot aber die erheblichen Kosten entgegen sowie der sicherlich auch erhöhte Zeitbedarf für die ausführliche Indexierungsarbeit. Hier kann die **elektronische Form** der Register eine Lösung bieten. Elektronische Register könnten, auch zeitversetzt, im Internet zur Verfügung gestellt werden. Es könnte auch zugleich mit dem Kauf eines Buches die Option auf Zusendung der Register in einem geläufigen Textverarbeitungssystem erworben werden, oder ein solches Register kann dem Buch sogleich als CD beigefügt werden. Ein elektronisches Register bietet gegenüber einer gedruckten Version stark erweiterte Zugangsmöglichkeiten.

Wenn das alphabetische und/oder das systematische Register in elektronischer Form zur Verfügung stehen, dann kann nach jedem beliebigen Wort gesucht werden oder auch nach einem Wortteil in einem Schlagwort. Es werden alle Schlagwörter gefunden, in denen die gesuchte Zeichenfolge auftritt, gleichgültig, an welcher Stelle im Schlagwort dies der Fall ist. Es wäre aber ein Trugschluss, zu glauben, dass man durch das blosse Umwandeln eines gedruckten Registers in eine elektronische Form alle Beschränkungen überwinden könnte, denen das gedruckte Register unterliegt. Erhalten bleiben stets die Nachteile, die ein Übermass an Präkombination, der Mangel an Grammatik und die oftmals erzwungene Lückenhaftigkeit bei den Verweisungen hervorrufen.[189] Ferner versprechen sich die meisten Nutzer von einer Computerrecherche als Ergebnis alles, was sie interessiert. Ein Retrievalresultat liefert jedoch immer nur das, wonach auch tatsächlich gesucht wurde.

Im folgenden Abschnitt beschäftigen wir uns mit der manuellen Produktion von Schlagwortregistern. In der Vergangenheit und in grossem Umfang auch noch in der Gegenwart werden Texte manuell, heute natürlich computerunterstützt, indexiert. Vielleicht kann das automatische Indexieren von den kognitiven Prozessen der Menschen bei der manuellen

[189] Fugmann 1999:175

Indexierung lernen. Das heisst nicht, dass diese Prozesse einfach in automatische Systeme kopiert werden sollten. Dennoch könnten Lösungen für gewisse Indexierungsprobleme der manuellen Vorgehensweisen für das automatische Indexieren von Nutzen sein. Das ist der Grund dafür, dass an dieser Stelle das Vorgehen der manuellen Indexierung beschrieben wird.

3.4 Das manuelle Vorgehen beim Erstellen eines Schlagwortregisters

Es gibt gewisse grundsätzliche Prinzipien, die für die Erstellung aller Indextypen gelten. Dennoch erfordert die Buchindexerstellung nach Ansicht vieler Fachleute eine spezielle Art des Indexierens. Die Indexierung eines Buches ist keine Arbeit für einen Amateur. Es ist eine anspruchsvolle Aufgabe, welche die Kenntnis eines Buchinhaltes, des Sachbereichs, der Fachterminologie sowie der Basisprozesse und der Methoden des Indexierens verlangt.[190] Aus diesem Grund führen geschulte und erfahrene Spezialisten, d.h. professionelle Indexierer, Buchindexierungen aus. Es gibt viele Richtlinien für das manuelle Indexieren von Büchern: Borko/Bernier 1978, Cleveland/Cleveland 1990, Fugmann 1999, Hutchins 1985, Lancaster 1991, Mulvany 1994, Rowley 1988, um nur ein paar zu nennen.

Zuweilen wird ein Buchindex vom Autor selbst geschrieben. In den letzten Jahren gab es etliche Diskussionen darüber, ob es klug ist, wenn Autoren Indexe kreieren. Einerseits wird betont, dass niemand mehr über den Inhalt eines Buches weiss als der Autor selbst und dass somit korrekte und sinnvolle Schlagwörter erwartet werden dürfen.[191] Andererseits wird darauf hingewiesen, dass ein Autor gewissermassen „vor lauter Bäumen den Wald nicht mehr sieht“ und deshalb die Leserperspektive nicht nachvollziehen kann. Ein von einem Autor erstellter Buchindex ist gemäss Cleveland/Cleveland und Rowley noch lange keine Garantie für einen guten und nützlichen Index.[192] Allerdings sind nicht einfach alle von professionellen Indexierern geschaffenen Register ohne Vorbehalt, denn ihnen fehlt oftmals die nötige Vertrautheit mit einem Fachgebiet, vor allem in sehr speziellen Bereichen.

Die physikalischen Hilfsmittel des professionellen Indexierers können erheblich variieren, je nach Grösse und Politik einer Indexagentur. Einige Agenturen benutzen vorgefertigte Formulare, und der Indexierungsprozess läuft auf ein Ausfüllen der Formulare hinaus, was natürlich auch auf dem Computer gemacht werden kann. Der Gebrauch von **Indexkarten** (*index cards*) ist speziell bei der Buchindexierung recht verbreitet. Bei dieser Methode werden die einzelnen Einträge auf je eine Karte geschrieben, die Karten alphabetisch sortiert sowie editiert, und anhand dieser Kartenanordnung wird ein Indexmanuskript getippt. Es stehen heute auch Softwareprodukte,

[190] Cleveland/Cleveland 1990:125; Lancaster 1998:104; Mulvany 1994:22; Moens 2000:55
[191] Mulvany 1994:22
[192] Cleveland/Cleveland 1990:125; Rowley 1988:23

welche die Kartenmethode auf dem Bildschirm simulieren, zur Verfügung.[193] Eine andere Methode ist die **Textdateimethode** (*text file method*). Das Verfahren ist sehr einfach: Die Einträge werden in eine Textdatei geschrieben, die Untereinträge kommen dabei indentiert unter ein Schlagwort. Werden neue Einträge hinzugefügt, so werden diese nach alphabetischer Ordnung platziert.[194]

Egal, welche Methode bei einer manuellen Indexierung angewendet wird, ein Indexierer sollte einige grundlegende Referenzwerkzeuge haben: ein aktuelles Standardwörterbuch und eventuell einen allgemeinen Thesaurus. Dazu kommen Wörterbücher, Handbücher und Thesauri aus dem Sachgebiet. Manchmal ist auch ein geografisches Lexikon oder ein Namenwörterbuch hilfreich. Wenn möglich, kann ein Indexierer ebenso einen oder mehrere Buchindexe aus dem gleichen Sachgebiet zur Unterstützung hinzuziehen. Wichtig ist die Kommunikation mit dem Herausgeber des Buches,[195] um gewisse Bestimmungen von vornherein festzulegen, z.B. Formatbestimmungen, sodass nachträgliche Überraschungen vermieden werden können.

Eine Indexierung involviert drei elementare Schritte. Der erste Schritt ist die konzeptuelle Analyse einer Textquelle und die Identifikation des Inhalts. In einem zweiten Schritt werden die ausgewählten und verallgemeinerten Inhalte in die Sprache der Textrepräsentation, in Indexterme, übersetzt. Der letzte Schritt besteht in der Nachbearbeitung der produzierten Einträge.

3.4.1 Inhaltsanalyse

Ist die Entscheidung für das Indexieren eines bestimmten Buches gefallen, werden vor dem Beginn der Inhaltsanalyse zuallererst die bibliografischen Daten der Textquelle erfasst. Die Erfassung derer ist recht einfach, das resultierende Format sollte aber konsistent sein.[196] Weiter muss, falls keine Anordnungen vorliegen, festgelegt werden, welche Teile des Buches der Indexierung unterzogen werden. Der eigentliche Text eines Buches wird selbstverständlich indexiert, relevante und periphere Informationen müssen dabei unterschieden werden. Fussnoten und Anmerkungen werden in die Indexierung mit einbezogen, wenn sie Material präsentieren, das im Text selbst nicht vorkommt. Bibliografische Angaben darin werden üblicherweise nicht indexiert. Es gibt eine allgemeine Regel, dass Illustrationen, Tabellen, Diagramme, Grafiken, Bilder, Fotografien und andere nicht textuelle Elemente in einem Buch nur indexiert werden, wenn die Darstellungen nicht auf derselben Buchseite stehen, wo das präsentierte Material im Text diskutiert wird. Liegen eine oder mehrere Seiten zwischen einer Illustration und den dazugehörigen Textpassagen, integriert man die Illustration in den Index.[197]

[193] Cleveland/Cleveland 1990:115; Mulvany 1994:240
[194] Mulvany 1994:243
[195] Cleveland/Cleveland 1990:127f.
[196] Cleveland/Cleveland 1990:104
[197] Mulvany 1994:45-48

Nach dem Erfassen der bibliografischen Daten und dem Festlegen der zu indexierenden Textstellen erfolgt die eigentliche **Inhaltsanalyse** (*content analysis*). Dabei wird zunächst das Thema des Buches ermittelt, indem ein Indexierer das Buch schnell durchliest im Sinne von Überfliegen (*scan*, *skim*) und ein allgemeines Gefühl für den Text und dessen Thema entwickelt, ohne an diesem Punkt wirklich zu indexieren, aber natürlich immer mit diesem Ziel im Hinterkopf. Er muss – ganz einfach ausgedrückt – entscheiden, worum es in einem Text geht. Die sogenannte *aboutness* eines Textes ist in den Textwörtern nicht immer explizit ausgedrückt und verlangt vom Indexersteller manchmal umfangreiches Sachgebietwissen. In einer zweiten Lesung wird nun das Buch gelesen und die Indexierung vorgenommen, indem substanziell abgehandelte Konzepte identifiziert und in Worten ausgedrückt werden, die eine Liste mit möglichen Schlagwörtern bilden. Als generelle Regel gilt: Jedes bedeutungsvolle Wort im Text wird in das Verzeichnis aufgenommen. Zu diesem Zeitpunkt ist es besser, zu viel in das Konzeptregister aufzunehmen als zu wenig. Es werden jetzt auch synonyme Terme sowie assoziative Begriffsbeziehungen aufgespürt und „siehe"- oder „siehe auch"-Verweise eingefügt. Falls ein Indexierer es für nötig befindet, wird er in einer weiteren Lesung das gesamte Buch nochmals mehr oder weniger detailliert durchgehen. Er hat nunmehr das vollständige Schlagwortvokabular vor Augen und kann so auf Lücken oder Unstimmigkeiten aufmerksam werden. Es werden dabei auch bestimmte Themen in spezifischere Ausdrücke übersetzt, andere durch allgemeinere Schlagwortkandidaten ersetzt.

Nicht jeder Text muss für die Essenzerkennung vollständig gelesen werden; stellenweises Lesen (*spot reading*) kann genügen, damit ein Indexierer versteht, welche Konzepte behandelt werden. Auf der anderen Seite müssen manche Texte komplett und vielleicht sogar mehr als einmal sehr genau gelesen werden. Eine Kombination von Überfliegen und Lesen wird im Indexierungsfachbereich befürwortet: Titel, Inhaltsverzeichnis, Kurzreferat oder Abstract (falls vorhanden), Einführung, Absätze am Anfang eines Abschnittes, Sätze am Beginn und Ende von Absätzen, Zusammenfassung, Schlussfolgerungen, Wörter oder Wortgruppen mit ungewöhnlichem Schriftbild (z.B. Unterstreichung, Kursivdruck), Illustrationen, Diagramme, Tabellen etc. sowie deren Erläuterungen sollen sorgfältig gelesen werden, sie sind sehr reich an potenziellen Schlagwörtern. Der Rest des Textes kann überflogen werden.[198]

Eine wirkungsvolle Inhaltsanalyse beinhaltet nicht nur, zu entscheiden, wovon ein Buch handelt, sondern auch zwischen der Essenz des Buches und nicht wirklichen Gegenstandsangaben zu differenzieren. Konzepte, die nur beiläufig erwähnt und nicht erläutert werden, finden keine Aufnahme in das Register. Ob ein bestimmter Gegenstand wert ist, indexiert zu werden oder nicht, hängt weit gehend von der Menge an Informationen ab, die dem Gegenstand gewidmet ist. Ein Indexierer sollte erahnen, welche Art von Fragestellungen ein Leser an das Register eines Buches stellen würde und welche Antworten hierbei wohl erwünscht sein könnten. Stellt er seine Überlegungen über die Interessen von bestimmten Benutzergruppen in den Vordergrund, strebt er eine **benutzerorientierte Indexierung** (*request-oriented indexing*) an. Richtet er seinen Fokus auf

[198] Moens 2000:56

den Text selbst, nennt man die Indexierung **dokumentorientiert** (*document-oriented*).[199] Bei der Indexierung von Büchern ist die Unterscheidung nicht so bedeutend; weder die eine noch die andere Art muss zwingend verfolgt werden.

3.4.2 Übersetzung in Indexterme

Als nächster Schritt folgt nach der Erkennung der Themen und der Zusammenstellung der Schlagwortkandidaten die Umwandlung derer in akzeptierbare Schlagwörter.[200] Diese Indexterme können – wie wir in Abschnitt 2.4.2 gesehen haben – extrahierte oder frei formulierte natürlichsprachliche Terme sein oder Terme aus einem Thesaurus, einer Klassifikation oder einer anderen verbindlichen Schlagwortliste. Wird ein kontrolliertes Vokabular verwendet, so werden die konzeptuellen Schlagwörter mit den Deskriptoren oder Notationen abgeglichen mit dem Ziel, die endgültigen Schlagwörter aus der vorgeschriebenen Indexsprache auszuwählen. Frei formulierte Schlagwörter kreiert man unter Beachtung grösstmöglicher Geläufigkeit und Eindeutigkeit, sie sollen prägnant und aussagekräftig sein. Die in den Texten vorkommenden Ausdrücke dienen für die Formulierung der Schlagwörter als Anregungen; es ist zusätzlich an die in einer Fachwelt üblichen Ausdrücke zu denken, auch wenn diese vom Autor des Textes nicht verwendet worden sind. Exhaustivität und Spezifität der selektierten Schlagwörter sind für eine spätere erfolgreiche Recherche Voraussetzung. Für Lancaster sind die Prinzipien der Exhaustivität und Spezifität die einzigen Regeln beim Zuteilen von Indextermen (obschon seiner Meinung nach sehr viele existieren, die besagen, was mit den Indextermen geschehen soll, wenn diese einmal ausgewählt worden sind). Jeder substanziell diskutierte Gegenstand, der für die Benutzer des Buches von Interesse sein könnte, wird so spezifisch wie möglich indexiert.[201] Ziel ist es, die Konsistenz der Indexierung und damit die Voraussehbarkeit und den Gebrauchswert bei der Benutzung zu erhöhen.

Teil der Übersetzung in Indexterme ist auch, jegliche Begriffsverwandtschaft (hierarchisch oder assoziativ) durch eine reziproke „siehe auch“-Verweisung auszudrücken, jegliche Wortverwandtschaft vom Typus der Synonymie durch eine gerichtete „siehe“-Verweisung in Richtung auf die Vorzugsbenennung. Man entnimmt die Verweisungen beider Typen aus dem kontrollierten Vokabular (falls vorhanden und angewendet) und trägt sie im Register vor den Untereinträgen bei den Schlagwörtern ein oder auch am Ende der Untereinträge. Die Meinungen sind geteilt, an welchem Platz die Verweisungen besser untergebracht sind.[202]

Obschon die Inhaltsanalyse und die Übersetzung in Indexterme zwei verschiedene Prozesse sind, können sie nicht immer klar getrennt werden und werden manchmal sogar simultan vollzogen.[203]

---

[199] Lancaster 1998:4
[200] Cleveland/Cleveland 1990:107; Moens 2000:57
[201] Lancaster 1998:31
[202] Fugmann 1999:165
[203] Lancaster 1998:14

3.4.3 Nachbearbeitung

Der letzte Schritt beim Indexieren eines Buches ist die Überprüfung dessen, was bisher gemacht wurde. Die Konsistenz der Einträge, die Richtigkeit der Querverweise, die Rechtschreibung und das Fehlen offensichtlicher Einträge werden kontrolliert. Mulvany empfiehlt, dass ein Register zuerst vom Indexierer und dann, wann immer möglich, vom Buchautor und auch von einem Redakteur editiert werden soll.[204]

Alle Schlagwörter werden im Hinblick auf ihre Funktion als primäre Zugangsstellen zum Index geprüft. Sie sollen sinnvoll sein, klar und kurz. Jedes einzelne Schlagwort aus einem Mehrwortschlagwort wird einmal als Ordnungswort an die erste Stelle gebracht, sodass der Suchende im Register unter einem jeden Schlagwort aus einer solchen Schlagwortkette den Zugang zu den betreffenden Textstellen findet. Schlagwörter mit vielen undifferenzierten Referenzstellen werden in Untereinträge aufgespaltet. Für die Untereinträge gelten die gleichen Überarbeitungskriterien wie für die Schlagwörter. Jeder Querverweis muss verifiziert werden. Blinde oder zirkuläre Verweisungen werden aufgesucht und geändert. Alle Fundstellen sollten korrekt sein; allerdings ist es nicht immer möglich, sie alle einzeln zu testen. Wenigstens sollte geprüft werden, ob sie vernünftig sind und sich eine Fundstelle beispielsweise nicht über Hunderte von Seiten erstreckt.

Alphabetisierung, Orthografie, Zeichensetzung, auch Format und Typografie werden aufmerksam kontrolliert; Konsistenz und Korrektheit sind dabei die wichtigsten Prüfsteine. Selbst wenn Computerwerkzeuge zu Hilfe genommen werden, ist eine nachträgliche manuelle Revision ratsam.[205] Kommen Sonderzeichen, Abkürzungen, Ländereigenheiten oder andere Besonderheiten im Register vor, wird eine präzise und knappe Vorbemerkung am Anfang des Verzeichnisses formuliert.

Um die adäquate Erfassung der Essenz zu prüfen, stellt sich der Überarbeiter Fragen wie: Decken die endgültigen Schlagwörter alle wichtigen Konzepte des Textes ab? Reflektieren die Schlagwörter die Gegenstände angemessen? Könnten die Einträge dazu benutzt werden, die Bedeutung und die Absicht des Buches zu rekonstruieren? Oder könnte anhand der Schlagwörter eine Synopsis geschrieben werden?[206] Es kann sich anlässlich dieser Fragestellungen auch die Notwendigkeit zu einer spezifischeren Indexierung ergeben. Neue Einträge werden sorgfältig eingefügt und können auch neue Querverweise zur Folge haben. Müssen wegen Platzmangels Einträge aus dem Register gestrichen werden, dann ist hierbei besondere Aufmerksamkeit angebracht, damit keine Lücken im

[204] Mulvany 1994:14
[205] Mulvany 1994:224
[206] Cleveland/Cleveland 1990:109

Netzwerk der Verweisungen auftreten.[207] Ein Register darf nicht einfach dadurch verkürzt werden, dass einzelne Einträge weggelassen werden. Präkombination eignet sich zur Inhaltserschliessung von Büchern besonders gut.

Immer wieder wird Konsistenz als Massstab für eine Qualitätssicherung beim Indexieren angegeben, sowohl bei der Inhaltsanalyse als auch beim Übersetzen in Indexterme sowie bei der Layoutfestlegung. Konsistenz wird durch verschiedene Faktoren beeinflusst: die Anzahl der zugeteilten Terme, die Verwendung eines kontrollierten oder freien Vokabulars, die Grösse und Spezifität des Vokabulars, die charakteristischen Merkmale eines Fachgebietes und dessen Terminologie, die Länge des Originaltextes, Erfahrung und Sachkenntnis des Indexierers, die zur Verfügung stehenden Werkzeuge und Hilfsmittel usw. Insbesondere die Verwendung eines verbindlichen Vokabulars wird als konsistenzfördernd angesehen. Lancaster macht einen Unterschied zwischen **vorschreibenden** (*prescriptive*) und **suggestierenden** (*suggestive*) verbindlichen Indexvokabularen, wobei Letzteres einem Indexierer eine gewisse Freiheit bei der Wahl der Indexterme überlässt, während ein vorschreibender Indexsprachenwortschatz keinerlei Freiheit gewährt. Es überrascht kaum, dass die vorschreibenden Vokabulare bessere Konsistenz produzieren.[208]

Nachdem ich das grundsätzliche Vorgehen der manuellen Registererstellung demonstriert habe, wende ich mich nun dem Aussehen eines Registers zu.

3.5 Form und Layout eines Schlagwortregisters

Die Gestaltung und das Aussehen eines Buchindex spielen für dessen Gebrauchswert eine wichtige Rolle. Das oberste Prinzip ist Konsistenz. Alle Entscheidungen bezüglich der Form und des Layouts eines Registers sollten dennoch auch mit Rücksicht auf die Bedürfnisse der Benutzer gefällt werden. Allzu oft treffen wir Indexe an, die auf ein paar wenige Buchseiten gequetscht worden sind und deren effektive Nutzung ein Zugeständnis zugunsten der Platzeinsparung darstellt. Bedauerlicherweise beeinträchtigen sich Benutzerfreundlichkeit und Platzeinsparung gegenseitig; beide stellen aber signifikante Aspekte bei der Gestaltung eines Buchregisters dar.

Ein Indexierer sollte sich Gedanken über den Umfang eines Index machen: Ist die Länge des Registers im Verhältnis zum gesamten Buch angebracht? Wie viele Schlagworteinträge soll der Buchindex haben? Wie viele Schlagwörter werden einem Thema zugestanden? Es existiert keine Formel für die „richtige“ Länge eines Buchindexes, aber es gibt gewisse Richtlinien. Viele verschiedene Umstände und Parameter bestimmten im Einzelfall die Grösse. Üblicherweise wird sie durch einen Kompromiss zwischen ökonomischen Beschränkungen und Indexierungsidealen

[207] Fugmann1999:164f.
[208] Lancaster1998:65

bestimmt. Cleveland/Cleveland schlagen pro hundert Buchseiten ungefähr fünf Registerseiten vor (also etwa 5% des Buchseitentotals).[209] Die Realisierung dieses Verhältnisses ist jedoch nicht immer möglich. Deshalb sollte ein Schlagwortregister eine Länge „von gesundem Menschenverstand" aufweisen und so viele Indexterme enthalten, dass das Thema im Buch sowie alle hauptsächlichen Benutzerzugriffe darauf vollständig abgedeckt sind.[210] In der folgenden Tabelle sind die von Mulvany empfohlenen Registerlängen von Büchern aus unterschiedlichen Bereichen in Prozent der Gesamtseitenzahl der Bücher ausgedrückt.

Buchtyp	Registerseiten in Prozent der Buchgesamtseitenzahl	Einträge pro Buchseite
Massenmarkt-Bücher leichte, nicht zu detaillierte Texte	2-5%	3-5
Allgemeine Fachliteratur Kochbücher, medizinische Texte, wissenschaftliche Bücher	7-8%	6-8
Technische Dokumentationen I allgemeine Benutzerhandbücher, einführende Handbücher, Ausbildungshandbücher	10%	8-10
Technische Dokumentationen II Gesetzbücher und Verordnungen, Service- und Reparaturmanuals, Systemhandbücher, Material für Spezialbereiche	15% +	10 +

Tabelle 2: Richtlinien für Registerlängen nach Mulvany[211]

Ob sich diese an sich nicht verbindlichen Zahlen bei der Untersuchung der Register im statistischen Teil der Arbeit bestätigen, wird sich in Abschnitt 5.1 zeigen.

Da Buchregister gedruckt werden, ist Platzeinsparung ein wichtiges Kriterium bei der Darstellung derselben. Es gibt verschiedene Wege, um in einem Register Platz zu sparen, der grundlegendste ist die Darstellung der Einträge. Die Gefahr dabei ist allerdings, dass die Lesefreundlichkeit geopfert und die Benutzung des Registers dadurch erschwert wird. Eine effektive Art, Platz zu sparen, ist die blockweise indentierte Anordnung der Untereinträge unterhalb eines Schlagwortes (siehe dazu auch Abschnitt 3.5.5 „Layout des Registers").

[209] Cleveland/Cleveland 1990:127
[210] Cleveland/Cleveland 1990:136
[211] Mulvany 1994:66

Grammatik 8, 39, 44
allgemeine oder universale 53; deskriptive 52; funktionale 51, 54; generative 54, 85, 333, 336, 374; normativ-präskriptive 52; traditionelle 53

Diese Gliederung eignet sich nur, wenn die Anzahl der Untereinträge sowie der Fundstellen nicht zu gross ist. Eine einfache eingerückte Anordnung trägt nicht zur Platzersparnis bei, ist aber im Gegensatz zur blockweisen Anordnung oder einer Gliederung ohne Indentierung sehr übersichtlich für den Benutzer, wie die folgenden Beispiele zeigen.

Grammatik 8, 39, 44
allgemeine oder universale 53
deskriptive 52
funktionale 51, 54
generative 54, 85, 333, 336, 374
normativ-präskriptive 52
traditionelle 53

Grammatik 8, 39, 44
Grammatik, allgemeine oder universale 53
Grammatik, deskriptive 52
Grammatik, funktionale 51, 54
Grammatik, generative 54, 85, 333, 336, 374
Grammatik, normativ-präskriptive 52
Grammatik, traditionelle 53

Abkürzungen und Akronyme sind heutzutage sehr beliebt, um Platz zu sparen. Solche Kurzformen sollten in einem Register nur verwendet werden, wenn die Benutzer sie mit Sicherheit verstehen werden oder wenn in der Vorbemerkung genügend Erklärungen stehen. Platzeinsparungen können auch durch die Anzahl der Leerschläge bei der Indentierung, die Anzahl der Spalten pro Seite, die Schriftgrösse usw. erreicht werden; Entscheidungen diesbezüglich verlangen jedoch stets das Miteinbeziehen der Nutzerperspektive.

3.5.1 Schlagwörter und Untereinträge

Die Schlagwörter, meistens Substantive oder von Adjektiven attribuierte Substantive, sollten von den Untereinträgen klar distinguiert werden können. Indentierung ist zu diesem Zweck sehr nutzbringend.

Es gibt an und für sich keine Begrenzung der Anzahl der Wörter eines Mehrwortterms. In einem der untersuchten Schlagwortverzeichnisse besteht das längste mehrgliedrige Schlagwort aus acht Tokens („Unbestimmtheit aller Theorien über die Natur, These der“[212]). Im Hinblick darauf, dass die meisten Schlagwortverzeichnisse in alphabetischer Reihenfolge angeordnet werden, sollte darauf geachtet werden, dass alle wichtigen Terme einmal am Anfang eines Eintrags stehen, sodass danach gesucht werden kann, auch wenn sich die Zahl der Schlagwörter dadurch vermehrt. Funktionswörter als Ordnungswörter sind zu vermeiden, z.B. „für das Netzwerk geeignete

[212] Stegmüller 1986:543

Hardware“ oder „andere wichtige Kopierbefehle“. Artikel, Präpositionen und Konjunktionen erscheinen nur am Anfang einer Zeile, wenn sie Teil einer Namensbezeichnung, z.B. Buchtitel, sind. Es herrscht in der Fachliteratur keine Einigkeit darin, welche Wortreihenfolge bei Mehrwortbenennungen, die aus einem Substantiv und einem Adjektiv bestehen, eingehalten werden sollte, ob z.B. „maschinelle Inhaltserschliessung“ oder „Inhaltserschliessung, maschinell“ vorzuziehen ist. Beide Varianten sind üblich, häufig tauchen sogar beide in demselben Register auf. Selbst wenn die Inversion ungewöhnlich klingt, hat sie den Vorteil, dass in einer alphabetischen Anordnung die verwandten Schlagwörter besser zusammengehalten werden. Ganze Sätze als Schlagwörter sind unzulässig, ausser es handelt sich um eine Namensbezeichnung oder um ein Zitat.[213]

Über die Verwendung von Singular- oder Pluralformen in Schlagwortregistern wurde schon viel diskutiert. Für Bonura und Cleveland/Cleveland spielt es keine Rolle, ob der Singular oder Plural gewählt wird, solange die Verwendung konsistent ist.[214] Kontrastierend dazu ist die Auffassung von Mulvany. Sie postuliert eine Aufführung der Pluralformen bei zählbaren Substantiven („Katzen“) und einen Gebrauch der Singularformen für unzählbare Substantive („Luft“). Tolerierbar sind für Mulvany auch Klammerschreibweisen: „Katze(n)“.[215] Indexierer bevorzugen üblicherweise entweder die eine oder die andere Form. Manchmal müssen aber beide Formen in einem Register verwendet werden, da die Singularform eines Wortes eine andere Bedeutung haben kann als die Pluralform. Zu vermeiden ist, dass die Singular- und Pluralformen eines Substantivs alphabetisch an verschiedenen Positionen eingeordnet werden (z.B. „Maus“, „Mäuse“). Klarheit und Zweckmässigkeit sollten das leitende Prinzip bei der Frage des Numerus sein, nicht einfach sture Konsistenz. Die Entscheidung ist auf jeden Fall schon vor dem Indexierungsprozess zu fällen.

Homographe in Registern erfordern zusätzliche Erklärungen für die Bedeutungsdifferenzierung. Häufig werden diese Erklärungen in Klammern gesetzt.

Basis (Mathematik) 190-196
Basis (Militär) 43, 47-51

Viele Indexierer glauben, dass die Indexierung von Eigennamen eine der einfachsten Aufgaben des Indexierens darstellt. Aber Eigennamen sind problematisch; ihre schier endlosen Variationen sollen selbst erfahrene Indexierer ins Schwitzen bringen.[216] Für Personennamen gilt die Regel: Der Nachname wird an erster Stelle aufgeführt, die Vornamen durch Komma abgetrennt dahinter:[217]

[213] Fugmann 1999:70
[214] Cleveland/Cleveland 1990:141; Bonura 1994:59
[215] Mulvany 1994:80f.
[216] Cleveland/Cleveland 1990:132
[217] Fugmann 1999:70; Mulvany 1994:79

Wolf, Christa 29, 35

Es ist meiner Meinung nach nicht notwendig, auf eine detailliertere Handhabung der Eigennamen einzugehen. Die vorzunehmenden Festlegungen sind sprachabhängig und können sich beispielsweise an den Empfehlungen der Fachliteratur orientieren. Es sollen hier nur ein paar Typen von Eigennamen erwähnt werden, deren Formen unter Umständen geklärt werden müssen: Namen berühmter Leute (Papst Leo XIII.), fremdsprachige Namen, Pseudonyme, Koautorschaften, historische Ereignisse, geografische Namen, Firmen, Institutionen, Regierungsämter, sehr verbreitete Personennamen (Hans Müller) usw. Für mehr Informationen zu diesem Thema sei auf Mulvany und Cleveland/Cleveland verwiesen.[218] Mulvany widmet den Eigennamen ein ganzes Kapitel, auch Cleveland/Cleveland erörtern die wichtigsten Aspekte der Eigennamenindexierung.

Bei manchen Registern werden bei den Untereinträgen die dazugehörigen Schlagwörter nicht wiederholt.

Einzelsprache
siehe Sprache
-: aktiver Einfluss auf Denken
-, Charakter der
-, Einfluss der

Es werden in vereinzelten Fällen sogar mehrere Bindestriche für Auslassungen verwendet. Dabei ist es nicht immer einfach, die Ellipsen richtig aufzufüllen.

Identität, Kriterium für
-, psycho-physische
- - , Kritik der
-, Substitutionsprinzip

Die Frage in diesem Beispiel ist, ob der zweite Untereintrag vollständig „Identität, Kriterium für, Kritik der“ oder „Identität, psycho-physische, Kritik der“ heisst. Auch wenn Auslassungen zur Vermeidung von Wiederholungen und für Platzeinsparungen getätigt werden, so wird ein Benutzer umso mehr Zeit aufwenden müssen, die Einträge richtig zu interpretieren. Ich persönlich ziehe beispielsweise Schlagwortregister ohne Auslassungen aus Übersichtlichkeitsgründen vor.

[218] Cleveland/Cleveland 1990:132-134; Mulvany 1994:152-182

3.5.2 Fundstellen

Die Referenzstellen werden meistens durch Kommas separiert, am Ende aller Referenzstellen zu einem Schlagwort steht kein Satzzeichen. Kommen in einem Register bibliografische Angaben vor, so ist sicherzustellen, dass eine klare Unterscheidung zwischen den Jahres-, Volumen-, Seitenzahlen etc. und den Fundstellen gemacht wird.[219] Um Platz zu sparen, können die Seitenzahlangaben gekürzt werden, indem einzelne Ziffern ausgelassen werden, z.B. „521-27“ anstelle von „521-527“. Zu beachten ist dabei, dass die verkürzten Angaben eindeutig interpretiert werden können. Für Kompressionsverfahren von Fundstellen gibt es verschiedene Richtlinien.[220] Die Platzeinsparung ist jedoch eher gering und lohnt sich nur, wenn wirklich ein Platzmangel vorherrscht.

Das Nachschlagen im einem Buch kann erleichtert werden, wenn diejenigen Passagen, in denen das Thema des Schlagwortes besonders ausführlich erörtert wird, im Register durch Fettdruck hervorgehoben werden.

> Arbitrarität 17, **20--21**, 92, 94, 97, 135

Bei diesem Eintrag erkennt ein Benutzer sofort, dass auf den Seiten 20 und 21 Definitionen oder substanzielle Erläuterungen zum Thema der Arbitrarität auffindbar sind.

Bezieht sich eine Fundstellenangabe auf ein Bild, eine Tabelle oder eine Fussnote, dann kann dies durch Anhängen von kennzeichnenden Buchstaben wie etwa „B“, „T“, „N“ verdeutlicht werden. Derartige Kennzeichnungen sollten in den Vorbemerkungen zum Register entsprechend erläutert werden.[221]

In einem elektronischen Index unterscheiden sich die Referenzstellen nicht von einem gedruckten Index. Der Unterschied besteht jedoch darin, dass ein Benutzer durch Anklicken der Fundstellen direkt auf die entsprechende Textstelle springen kann. Ein zusätzlicher Aufwand wird erforderlich: In einem elektronischen Schlagwortregister müssen alle Fundstellen mit den Sprungzielen im Text verlinkt werden. Auch ein Zurückspringen durch Anklicken des Ziels kann bei einer Recherche hilfreich sein.

[219] Cleveland/Cleveland 1990:140
[220] Mulvany 1994:87f.
[221] Fugmann 1999:159

3.5.3 Querverweise

Die formalen Ausgestaltungen und die Platzierung der Querverweise variieren stark. Einige Beispiele werden im vorliegenden Abschnitt präsentiert.

Fundstellen werden bei Querverweisen grundsätzlich nicht angegeben.[222] Der Platzbedarf wäre untragbar gross, wenn man die Referenzstellen zu einem Begriff unter jedem Synonym, welches ein Autor benutzt hat oder welches in der Fachliteratur gebräuchlich ist, vollständig aufzählen würde. Alle Fundstellenangaben zu einem betreffenden Schlagwort werden an einer bevorzugten Stelle zusammengetragen. Im Gegensatz zu den Untereinträgen werden sie meistens auch nicht in alphabetischer Reihenfolge angeordnet. Der Benutzer muss vorsorglich sowieso sämtliche Verweisungen überprüfen.

Für einen Suchenden ist es von Bedeutung, dass er innerhalb eines Eintrags die Querverweise, die vom Schlagwort ausgehen, und die Querverweise, die von einem Untereintrag ausgehen, deutlich unterscheiden kann. Sie müssen also entsprechend positioniert sein. Ein von einem Schlagwort ausgehender Verweis steht auf der gleichen Zeile wie das Schlagwort oder am Anfang bzw. am Ende aller Untereinträge als eigener Untereintrag, während ein von einem Untereintrag ausgehender Verweis auf der Zeile des Untereintrags oder auf einer eigenen Zeile darunter steht, dann jedoch nochmals indentiert. Verweise, die von Untereinträgen ausgehen, sollten nach Ansicht von Mulvany grundsätzlich vermieden werden, denn die Begriffsverbindungen in einem Index sollten auf der Ebene der Schlagwörter stattfinden.[223]

Da die Querverweise nicht den Zugang zu Informationen in einem Text bezwecken, sondern der Navigation innerhalb eines Registers dienen und verschiedene Indexterme miteinander verbinden, werden „*siehe*“ und „*siehe auch*“ sehr oft durch Kursivdruck gekennzeichnet. Zuweilen werden sie auch in Grossbuchstaben geschrieben („SIEHE“), abgekürzt („s.“; „s.a.“) oder in Klammern gesetzt. Diese Markierungen sind fakultativ. Wenn ein Querverweis zu mehr als nur einem Schlagwort hinweist, werden diese nicht mit einem Komma, sondern mit einem Semikolon voneinander getrennt („Buch, *siehe auch* Referenzbuch; Handelsbuch“), um die Verschiedenheit der Sprungziele zu betonen. Wird das Verweisziel kursiv gedruckt, so bedeutet das, dass ein Benutzer kein solches Schlagwort vorfinden wird, sondern dass es sich um einen Verweis auf eine Klasse von Schlagwörtern handelt. Er wird beispielsweise bei einem Verweis „Südostasien *siehe einzelne Ländernamen*“ alle Namen der südostasiatischen Länder durchsuchen müssen.

Wie auch immer die Form der Quervereise aussieht und wo auch immer sie positioniert sind, so gilt bei den Querverweisen wie auch beim gesamten Schlagwortregister, dass Konsistenz, Uniformität sowie Benutzerfreundlichkeit vorrangige Bedeutung besitzen.

---

[222] Fugmann 1999:68; Mulvany 1994:108
[223] Mulvany 1994:191

3.5.4 Anordnung der Einträge

Die Anordnung der Einträge referiert auf die Reihenfolge, in welcher die Schlagwörter präsentiert werden. In den meisten Schlagwortregistern ist diese Anordnung alphabetisch. Die alphabetische Sortierung kann **Wort für Wort** (*word-by-word order*) oder **Buchstabe für Buchstabe** (*letter-by-letter order*) vorgenommen werden. Wenn Wort für Wort alphabetisiert wird, werden zuerst die Zeichen des ersten Wortes sortiert. Das zweite Wort und alle nachfolgenden werden nur geordnet, wenn zwei oder mehr Einträge mit dem gleichen Wort beginnen. Das Leerzeichen kommt vor den Buchstaben, Kommas werden ignoriert. Ein Beispiel:

New Hampshire
New Jersey
New York
Newark
Newton
Newton, Isaac

In einer strikten Buchstabe-für-Buchstabe-Alphabetisierung werden nur die alphanumerischen Zeichen geordnet und alle anderen ignoriert, als ob alle Wörter eines Schlagwortes zusammengeschrieben wären:

Newark
New Hampshire
New Jersey
Newton
Newton, Isaac
New York

Es existieren auch Varianten dieser Sortierung. Beispielsweise wird die Alphabetisierung bei Kommas, die der Nachstellung von Termbestandteilen dienen, unterbrochen. Alle Sonderzeichen (Bindestriche, Schrägstriche, Doppelpunkte usw.) werden ignoriert.

Zahlen, Sonderzeichen und Satzzeichen machen die alphabetische Ordnung zu einem auf den ersten Blick vielleicht doch nicht so einfachen Prozess. Die alphabetischen und die nichtalphabetischen Elemente werden deshalb alle zu **Zeichen** (*characters*) zusammengefasst. Diese Zeichen umfassen Buchstaben, Zahlen, Sonderzeichen und auch Leerzeichen. Innerhalb der Zeichen gilt meistens die folgende Rangfolge: Leerzeichen, Sonderzeichen, Zahlen, Buchstaben.[224] Auch römische Zahlen,

[224] Mulvany 1994:113

fremdsprachige Buchstaben, Anführungs- und Schlusszeichen können Bestandteil eines Schlagwortes sein. Mulvany schlägt zu deren Behandlung vor, Apostrophe sowie Anführungs- und Schlusszeichen zu ignorieren, die römischen und arabischen Ziffern anhand ihrer numerischen Werte zu ordnen (aufsteigend), fremdsprachige Buchstaben (à, é, ñ, α, β etc.) wie die „normalen" Buchstaben oder ihre Entsprechungen in der eigenen Sprache zu sortieren. Die Umlaute im Deutschen können wie *ae, oe, eu* oder wie *a, o* und *u* behandelt werden.

Je nachdem, nach welcher Art die Schlagwörter alphabetisch sortiert werden, Wort für Wort oder Buchstabe für Buchstabe, werden auch die Untereinträge auf gleiche Weise geordnet. Die Wahl der Alphabetisierung ist einerlei, sie sollte sich im gesamten Register konsistent durchziehen und bei Bedarf in der Vorbemerkung zum Register erklärt werden. Eine Sortierung der Untereinträge bringt gemäss Fugmann wegen der Unvorhersehbarkeit des Wortlautes nur wenig Nutzen.[225] Aus meiner Erfahrung wird jedoch der Umgang mit Registern gerade auch durch die alphabetische Sortierung der Untereinträge erleichtert. Eine nichtalphabetische Anordnung auch der Schlagwörter findet sich offenkundig bei den systematischen Registern.[226]

3.5.5 Layout des Registers

In diesem Abschnitt werden einige Punkte bezüglich der Darstellung von Schlagwortverzeichnissen sowie der Typografie und Seitengestaltung angesprochen.

Ein Buchindex wird in derselben Schriftart gedruckt wie der Buchtext selbst, meist zwei Schriftgrade (*points*) kleiner. Wenn also die Schrift eines Buches *Times Roman* mit Schriftgrad 10 ist, so wird auch der dazugehörige Index in *Times Roman*, jedoch mit einem Schriftgrad 8, gedruckt. Ein kleinerer Schriftgrad rechtfertigt sich dadurch, dass ein Register nicht wie ein Text gelesen, sondern mit dem Wissen um die alphabetische Sortierung überflogen wird. Meistens werden zwei, bei sehr grossen Buchseiten auch drei oder sogar vier Spalten pro Seite aufgeführt. Das Ziel der Seitengestaltung ist, möglichst viele Einträge auf eine Seite zu bringen und sicherzustellen, dass die Einträge dennoch einfach abgesucht werden können. Jeder Buchindex sollte eine Überschrift haben. Existiert ein einziger Index, so ist die Überschrift „Index" oder „Register" ausreichend. Sind mehrere Verzeichnisse vorhanden, sollte jedes einen klaren Titel erhalten, z.B. „Namenverzeichnis", „Sachregister".

Um die Schlagwörter in einem Buchverzeichnis von den Untereinträgen deutlich unterscheiden zu können, wird eine Einrückung oder **Indentierung** (*indention, indented style*) der Untereinträge vorgenommen. Die Indentierung visualisiert die Hierarchie der Einträge. Ein Spezialfall der

[225] Fugmann 1999:158
[226] Mulvany 1994:121

Einrückung ist die **blockweise Indentierung** (*run-in style*).[227] Der Unterschied zwischen den beiden Stilen liegt in der Formatierung der Untereinträge. Bei beiden Formaten werden die Schlagwörter nicht indentiert, sämtliche Untereinträge und über das Zeilenende hinauslaufende Zeilen werden eingerückt präsentiert, wie die folgenden Beispiele zeigen.

Indentierung:

Grammatik 8, 39, 44
  allgemeine oder universale 53
  deskriptive 52
  funktionale 51, 54
  generative 54, 85, 333, 336, 374
  normativ-präskriptive 52
  traditionelle 53

blockweise Indentierung:

Grammatik 8, 39, 44
  allgemeine oder universale 53; deskriptive 52; funktionale 51, 54; generative 54, 85, 333, 336, 374; normativ-präskriptive 52; traditionelle 53

Bei einem eingerückt dargestellten Register steht jeder Untereintrag auf einer neuen Zeile mit einem bestimmten Mass an Indentierung. Dieser Layouttyp wird im Englischen auch *setout*, *hierarchical*, *outline* oder *line-by-line style* genannt. Das Mass der Indentierung soll in einem sinnvollen Verhältnis zur Schriftgrösse stehen[228], sodass die Untereinträge nicht zu weit links oder rechts im Verhältnis zum Schlagwort stehen. Wenn in einem Eintrag mehrere Einbettungsstufen existieren, z.B. Untereinträge von Untereinträgen von Untereinträgen, sollte auf jeden Fall eine Indentierung angewendet werden. Über ein Zeilenende hinauslaufende Einträge werden von den Untereinträgen gesondert. Allerdings wird häufig für Untereinträge und übers Zeilenende hinausgehende Schlagwörter das gleiche Mass an Indentierung benutzt. Die Methode funktioniert meiner Auffassung nach recht gut. Für Mulvany gibt es nur einen einzigen Grund, eine blockweise Einrückung vorzunehmen. Der Grund ist Platzeinsparung (vergleiche Abschnitt 3.5).[229] Denn bei der blockweisen Indentierung ist das Absuchen und Lokalisieren der Untereinträge für den Benutzer nicht gerade einfach. Sie eignet sich entsprechend für Register, die aus Schlagwörtern und nur einer darunter liegenden Ebene von Untereinträgen bestehen.

[227] Mulvany 1994:184-187
[228] Mulvany 1994:185
[229] Mulvany 1994:186

Typografische Kennzeichnungen können die Nützlichkeit eines Registers verbessern. Dennoch gilt die Regel: Je einfacher und einförmiger die Typografie, desto besser.[230] Denn Buchindexe tendieren dazu, dicht und gedrängt dargestellt zu sein. Obschon diese Richtlinie meinem eigenen Empfinden entspricht, steht sie in Kontrast zu Fugmanns Aussage, dass ein Register auch durch unterschiedliche Buchstabengrösse und durch Fettdruck bei den Schlagwörtern sehr übersichtlich gestaltet werden könne.[231] Ein kurzer Blick auf das Schlagwortregister von Fugmann genügt, um diesem die Übersichtlichkeit abzuerkennen; in vielen Fällen ist es für das Nachschlagen von Fachtermini sogar sinnvoller, das zum Buch gehörige Glossar zur Hand zu nehmen.

Ein vollständiger Verzicht auf typografische Kennzeichnungen ist trotz allem nicht angebracht. Der Kursivdruck bei Querverweisen ist weit verbreitet. Ebenso leistet der Fettdruck bei Fundstellen gute Dienste; ein Benutzer weiss, dass er dort Definitionen und substanzielle Diskussionen finden kann, ohne dass dabei eine zusätzliche Zeile im Register verbraucht worden ist. Speziell bei Softwarehandbüchern hilft die Verwendung einer anderen Schriftart (üblicherweise *Courier*), um Befehle, die genau so eingetippt werden, von Konzepten zu unterscheiden. Existieren in einem Schlagwortregister ausserordentlich viele Untereinträge zu fast allen Schlagwörtern, so können in diesem Fall die Schlagwörter fett gedruckt werden.

In einem Schlagwortregister sollten Zeilenumbrüche, die unansehnlich und verwirrend sind, vermieden werden. Sogenannte *bad breaks* vermindern die Nutzungsmöglichkeiten eines Registers. Das Problem dabei ist, dass sie erst zum Vorschein kommen, wenn Formatierung und Layout ausgeführt werden. Meiner Ansicht nach sollte ein Buchindex z.B. in XML annotiert sein (insbesondere auch im Hinblick auf elektronische Ausgaben). Die Formatierung und Layoutgestaltung bleibt dadurch flexibel und kann jederzeit geändert und angepasst werden. Unschöne Zeilen- oder Seitenumbrüche werden auf einfache Art unterdrückt.

Wenn sehr umfangreiche und dichte Buchregister erstellt werden, sollte das Anfügen des ersten Schlagwortes einer Seite ganz oben auf der Seite wie in einem Wörterbuch in Betracht gezogen werden. Bei einem alphabetischen Register kann auch der Wechsel im Anfangsbuchstaben bei den Schlagwörtern optisch verdeutlicht werden, indem eine Zwischenzeile und der neue Buchstabe als Überschrift eingefügt werden. Im systematischen Register ist es besonders hilfreich, wenn am Anfang einer jeden neuen Druckseite ein Überblick darüber gegeben wird, an welchem Platz in der Hierarchie man sich mit dem ersten Schlagwort auf dieser Seite befindet. Idealerweise wird jeder Seiten- oder Spaltenumbruch innerhalb eines Eintrags am Ort der Fortsetzung mit „Fortsetzung“ oder „fortgesetzt“ eingeleitet.

Als Zusammenfassung dieses Abschnittes könnte man sagen, dass der Form der Einträge sowie dem Layout genügend Aufmerksamkeit zukommen und das Design möglichst vor der eigentlichen Indexierung festlegt werden sollte. Dann kann der Indexierer bereits beim Formulieren der

[230] Mulvany 1994:194
[231] Fugmann 1999:173

Indexterme entsprechende Markierungen vornehmen (z.B. die Angabe von „T“ für Tabelleninformationen). Oberste Priorität geniessen abermals Konsistenz, Uniformität und Benutzerfreundlichkeit.

Damit haben wir eine kleine Auswahl an Ansprüchen an ein Schlagwortregister vorweggenommen, und wir wollen im folgenden Abschnitt auf die generellen Anforderungen an ein Buchregister genauer eingehen.

3.6 Die Anforderungen an ein Schlagwortregister

Die Anforderungen, denen ein Schlagwortregister genügen sollte, sind je nach verfolgtem Ziel von Fall zu Fall verschieden. Eine Indexierung kann die Absicht haben, exhaustiv oder selektiv zu sein. **Exhaustive Indexierung** (*exhaustive indexing*) impliziert die Verwendung einer ausreichenden Anzahl von Indextermen, um das Thema eines Dokuments möglichst vollständig abzudecken. Die Indexierung ist ausgesprochen erschöpfend und spezifisch; sie erreicht gute Retrievalergebnisse. Eine **selektive Indexierung** (*selective indexing*) meint die Verwendung einer kleineren Anzahl von Indextermen, um die zentralen Themen eines Dokuments zu repräsentieren. Das Suchergebnis ist zwar weniger spezifisch als beim exhaustiven Indexieren, dafür aber die Indexierung schneller und wirtschaftlicher durchzuführen. Vielfach wird der Begriff **tiefes Indexieren** gebraucht, um sich auf das exhaustive Indexieren zu beziehen, **breites Indexieren** meint die selektive Indexierung.[232] Man kann also aufgrund der unterschiedlichen Zielsetzungen keine allgemeingültigen Anforderungen formulieren, dennoch sollten gewisse Mindestanforderungen von jedem Buchregister erfüllt werden:

Korrektheit: Ein Schlagwortregister sollte keine Fehler enthalten, keine Zuweisung von falschen Schlagwörtern, keine Verwendung von Wörtern, welche die Bedeutung eines Schlagwortes verändern, keine Orthografiefehler, keine inkonsistente Zeichensetzung, keine falschen Referenzstellenangaben, keine blinden oder zirkulären Querverweise usw.

Spezifität: Bei jedem Schlagwortregister ist eine möglichst hohe Spezifität der Schlagwörter anzustreben. Im Gegensatz zu den Information Retrieval-Systemen besteht beim Buch jedoch nicht die Ungewissheit über die zukünftige Grösse des Systems, und man kann sich mit demjenigen Grad an Spezifität begnügen, welcher im Text selbst vorherrscht.

Exhaustivität: Die Abdeckung eines Registers soll so breit wie möglich sein und keine Erfassungslücken aufweisen. Die Exhaustivität (wie auch die Spezifität) wird im Hinblick auf den

[232] Kaiser 1996:3; Lancaster 1998:23f.

Fachbereich, den möglichen Leser sowie die Länge des Schlagwortregisters determiniert. Dabei sollte man nicht vergessen, dass hohe Abdeckung und hohe Spezifität die Anzahl der Indexterme wesentlich erhöht und die Gesamtlänge des Index zunimmt.[233]

Konsistenz und **Uniformität**: Auch wenn die Buchautoren dazu tendieren, in ihrer Wortwahl inkonsistent zu sein, so muss ein Register sowohl in Bezug auf die inhaltlichen Aspekte der Indexterme wie auch vor allem auf die formale Ausgestaltung konsistent sein.

Benutzerfreundlichkeit: Ein Leser soll möglichst einfach zu den Informationen im Text hingeführt werden. Alle Entscheidungen über das Design eines Schlagwortregisters werden deshalb unter Berücksichtigung der Benutzerperspektive getroffen. Selbst der einfachste und kürzeste Buchindex sollte Unterbegriffe haben, um eine grosse Anzahl von Fundstellen bei einem Schlagwort zu differenzieren. Das Ausmass und die Komplexität der Schlagwörter und Untereinträge hängen von der Länge des Buches, der Grösse des Buchindex und dem Sachgebiet ab. Es ist erforderlich, dass aus einem Register hervorgeht, ob sich ein Thema über eine Seite oder über mehrere Seiten hinweg erstreckt. Die Kennzeichnungen der Fundstellen mit „f." oder „ff." sind mindestens erforderlich, besser natürlich Bindestrich-Angaben. Querverweise werden dort angebracht, wo sie aus Sicht der Suchenden für nützlich erscheinen.

Auch eine Vorbemerkung sollte im Register bei Abweichungen von der Norm nicht fehlen. Diese gibt beispielsweise an, welche Buchteile nicht indexiert worden sind oder nach welchem Prinzip die alphabetische Reihenfolge unter den Schlagwörtern hergestellt worden ist, insbesondere beim Auftreten von Umlauten, Sonderzeichen und Leerzeichen in Mehrworttermen.

Angemessene Länge: Ein Register sollte eine angemessene Länge im Verhältnis zu dem gesamten Buchtext und bezüglich des behandelten Sachgebietes aufweisen. Eine Abgleitung ins Pedantische ist ebenso zu verhindern wie eine Beschränkung auf das Minimum, z.B. aus Gründen der Platzeinsparung.

Erkennen der Essenz: Ein Register soll die Essenz eines Buches widerspiegeln. Auch dabei spielt die Perspektive des Nutzerkreises eine Rolle, d.h., dass zum Indexierungszeitpunkt zu erahnen versucht wird, welche Art von Fragestellungen an das Buchregister herangetragen werden. Peripheres ist von der Indexierung auszuschliessen, es soll also nicht auf Passagen verwiesen werden, an welchen das Fragethema nur beiläufig erwähnt, aber nicht ausführlich erörtert wird. Ebenso werden Bemerkungen, die zur Einführung in oder Erinnerung an ein Thema dienen, sowie allseits bekanntes Grundwissen nicht in das Verzeichnis aufgenommen. Dennoch muss eine Suche nach allgemeinen Konzepten ausführbar sein.

[233] Cleveand/Cleveland 1990:128

Lexikalisierung: Das Schlagwortregister stellt für die paraphrasierenden Ausdrucksweisen der Buchautoren Begriffsbenennungen zur Verfügung. Die Terminologie wird so gewählt, dass ein Benutzer im Register üblicherweise die Terme vorfindet, die er auch erwartet.

Bedeutungsfestlegung: Um Mehrdeutigkeiten aufzulösen und die Bedeutung von Fachausdrücken zu klären, legt ein Register die genauen Bedeutungen der Schlagwörter und Untereinträge fest. Denn Mehrdeutigkeiten können bei einer Recherche leicht Ursache von viel Ballast sein.

Terminologie- und **Begriffskontrolle**: Synomymie, Polysemie, Homonymie, Hyponymie und Hyperonymie sowie assoziative Begriffsrelationen werden in einem Buchregister durch Querverweise erfasst.

Ellipsenauffüllung: Ein Register sollte die Lücken und Auslassungen eines Buchtextes (z.B. zur Vermeidung von Wiederholungen) auffüllen, damit eine Suche nach diesen zusätzlichen Schlagwörtern möglich wird.

Präkombination: In einem gedruckten Register besteht eine ausgeprägte Notwendigkeit für Präkombination, um höhere Spezifität zu erreichen und halbwegs ballastarm recherchieren zu können. Die Nutzen und Nachteile von Präkombination sind bei einem Buchindex anders zu beurteilen als z.B. bei einer Datenbank, da im Register kaum die Möglichkeit einer Indexsprachengrammatik zur Verfügung steht und der Umfang des Registers nicht in einem unerträglichen Ausmass anwachsen wird.

Begriffsanalyse: Im Buchregister ist die Begriffsanalyse obligatorisch, sodass auch nach Begriffen, die Bestandeile von anderen merkmalsreicheren Begriffen sind, gesucht werden kann. Denn bei gedruckten Registern existiert keine Möglichkeit der Postkombination, die Schlagwörter können also nicht nachträglich verändert werden.

Voraussehbarkeit und **Rekonstruierbarkeit**: Die Schlagwörter eines Buchregisters müssen auch zu einem zukünftigen Zeitpunkt vom Benutzer rekonstruierbar sein. Er stellt sich bei einer Suche ein Schlagwort vor, unter welchem das Thema der Fragestellung auffindbar sein könnte. Die Recherche bleibt erfolglos, wenn dem Suchenden kein passendes Schlagwort einfällt und wenn er auch durch Verweisungen nicht auf ein solches Schlagwort hingewiesen wird. Die Lexikalisierung von Paraphrasen ist eine wichtige Massnahme zum Erzielen von Voraussehbarkeit.

Dies sind die Anforderungen, denen ein Schlagwortregister gerecht zu werden versuchen sollte. Das Problem dabei ist, dass sich die Forderungen zum Teil gegenseitig beeinträchtigen, sodass in jedem einzelnen Fall je nach Sachgebiet, Benutzerkreis oder anderen vorgeschriebenen Richtlinien die anzustrebenden Ziele gegeneinander abgewogen werden müssen. In einer ausgedehnten Studie von Hauck wurde festgestellt, dass nicht ein einziges von zehn untersuchten, modernen Fachbüchern im

Bereich der Datenverarbeitung die Minimalanforderungen erfüllte, die man an ein Buchregister stellen darf.[234] Eine der Ursachen für diese Sachlage dürfte darin liegen, dass aus wirtschaftlichen Überlegungen heraus heute weit verbreitet vollautomatische Indexierungsverfahren ohne manuelle Nachprüfung angewendet werden. Trotzdem hält sich der durch mangelhafte Indexierung verursachte Schaden beim Buchregister halbwegs in Grenzen, weil er auf das einzelne Buch beschränkt bleibt und weil ein Benutzer in dringenden Fällen durch stundenlanges Suchen meistens doch noch fündig werden kann.[235]

Mit genau festgelegten Regeln können die Anforderungen an ein Schlagwortregister besser erfüllt werden. Diesbezüglich sind zwei Bemerkungen zu machen. Erstens ist eine sture Orientierung an Regeln nicht immer eine Garantie für Qualität und Konsistenz; zweitens ist es gemäss Cleveland/Cleveland unmöglich, eine Liste mit Indexierungsregeln zusammenzustellen, welche jede mögliche Situation abdecken, weil zu viele Variablen bei jedem einzelnen Indexierungsprozess involviert sind. Wenn es möglich wäre, das Indexieren auf ein paar Regeln zu reduzieren, so wäre nach ihrer Meinung eine Automatisierung schon lange realisiert worden.[236]

Das Stichwort *automatische Indexierung* veranlasst uns dazu, auf die weiter oben formulierte Frage nach dem automatischen Erstellen von Schlagwortregistern zurückzukommen. Wie können wir die in diesem Kapitel gemachten Ausführungen für eine Automatisierung nutzen? Alle formalen Aspekte eines Registers, z.B. die Formatierung und die Layoutgestaltung, können insbesondere bei der Verwendung einer Markup-Sprache recht einfach automatisiert werden. Als ausgesprochen schwierig hingegen entpuppen sich die intellektuellen Prozesse der Inhaltsanalyse. Im zweiten, statistischen Teil der Arbeit werden wir deshalb eine Methode testen, die nicht versucht, das manuelle Vorgehen zu simulieren, sondern für das automatische Indexieren von Texten auf andere Techniken ausweicht.

Damit schliesse ich das Kapitel *Schlagwortregister* ab, und wir beschäftigen uns im folgenden letzten Kapitel des theoretischen Teils im Anschluss an die soeben angedeuteten Äusserungen bezüglich der Qualität von Buchregistern generell mit der Bewertung der Qualität von Indexen.

[234] Hauck 1997
[235] Fugmann 1999:132
[236] Cleveland/Cleveland 1990:100

4 Evaluation eines Index

Bei einem Vergleich von manuell und automatisch erstellten Indexen stellte sich heraus, dass sich in je einem Drittel die manuelle Indexierung einerseits und die automatische Indexierung andererseits als überlegen zeigte. Im verbleibenden Drittel schnitten beide Indexierungsverfahren gleichermassen gut ab.[237] Es stellt sich diesbezüglich die Frage, wie solche Vergleiche und Bewertungen angestellt werden. Um die Beantwortung genau dieser Frage geht es in diesem Kapitel. Wir werden uns mit den Charakteristika von sogenannt „guten" Indexen, den Faktoren, welche die Indexierungsqualität beeinflussen können, sowie mit den wichtigsten Methoden und Massen zur Evaluation von Indexen beschäftigen. Zunächst aber möchte ich auf ein paar Aspekte hinweisen, die man sich beim Studium jeder Indexierungsbewertung vor Augen halten sollte.

Indexieren ist stets nur Mittel zum Zweck. Dies hat zur Folge, dass eine Definition von „guter Indexierung" notwendigerweise unbefriedigend bleiben muss, solange nicht der aktuelle Zweck genau spezifiziert ist. Eine Indexierung für ein Retrieval (z.B. Forschungsrecherche) und eine Indexierung für ein gedrucktes Register oder für Spezialauswertungen (z.B. die Erstellung spezieller Reporte) können nicht nach denselben Kriterien bewertet werden. Somit wird jede Evaluation schwierig und bis zu einem gewissen Grad je nach Festlegung der Methode und Beurteilung der Suchergebnisse auch subjektiv. Die Qualität eines Index hängt in grossem Ausmass vom Benutzer ab. Er muss verstehen, was er will und wonach er eigentlich sucht. Die Qualität der Suchstrategie (z.B. Erfahrung, Scharfsinn etc.) hat einen starken Einfluss auf die Performanz eines Informationssystems.

Das Ziel eines jeden Index ist die Unterstützung des Benutzers bei der Suche nach Informationen in Dokumenten innerhalb einer angemessenen Zeitspanne und mit einem vertretbaren Aufwand auf Seiten des Benutzers. Das Wesen der wieder zu findenden Informationen bzw. das Wesen von Information allgemein macht jede Indexevaluation zu einer diffizilen Angelegenheit, weil Information keine physikalische Grösse oder Ware im materiellen Sinn ist. Information kann man nicht besichtigen, zur Probe benutzen und bei Nichtgefallen wieder zurückgeben. Fordert man Information an, dann ist man im Voraus bereit, für diesen Auftrag zu bezahlen, obwohl man noch nicht weiss, ob man das Bestellte auch wirklich bekommt. Zuweilen ist sogar die Nichtlieferung das eigentlich Erwünschte (wenn z.B. keine Neuveröffentlichung zu einem bestimmten Thema vorliegt). Ein Mangel an Information ist nicht so offenkundig wie der Mangel an einem Fahrzeug, einem Kühlschrank usw. Man weiss nie, was man alles nicht weiss. Auch hat ein Mangel an Information nicht derart akute und spürbare Folgen wie der Mangel an einem materiellen Gut (Ausfall des Fernsehers, fehlendes Auto oder fehlender Kühlschrank). Die Mehrfachlieferung der gleichen Information bedeutet keinen Mehrbesitz, sondern zuweilen sogar Belästigung. Zwei Bücher oder zwei Zeitungen genau der gleichen Art haben nicht doppelte Information zur Folge;

[237] Knorz 1994:28

zwei Kühlschänke hingegen ergeben die doppelte Kühlkapazität. Es liegt im Wesen von Information, dass sie für den Empfänger stets einen gewissen Wert hat. Dieser Wert ist jedoch für jeden Menschen verschieden, je nachdem, in welcher Lage er sich befindet, wie seine augenblickliche Interessenlage ist, welche einschlägigen Kenntnisse er zum Thema der Information bereits besitzt, ob er die Sprache beherrscht, in welcher die Information mitgeteilt wird, usw. So lässt sich für Information kaum ein objektiv gültiger Wert angeben. Für Fugmann ist der Nutzen von treffender Information nicht sicher ermittelbar und schon gar nicht quantifizierbar.[238] Dennoch wollen wir die Möglichkeit haben, unterschiedliche Indexierungen und Recherchenergebnisse vergleichen und gegeneinander abwägen zu können, auch wenn wir dabei eine Quantifizierung von Informationen vornehmen, indem wir sagen, dass eine gefundene Information relevant ist oder nicht. Exakt an diesem Punkt setzten die meisten hier erörterten Evaluationsmethoden an.

4.1 Merkmale eines „guten“ Index

Die Definition eines „guten“ Index ist nach Ansicht von Cleveland/Cleveland einfach: Ein „guter“ Index ist ein Index, der einen Benutzer genau zu den Informationen führt, die benötigt werden, ohne auf grössere Hürden zu stossen und ohne auf irrelevantes Material zu verweisen.[239] Korrespondierend ist „gute“ Indexierung bei Lancaster eine Indexierung, die das Wiederfinden von Einheiten in einer Datenbank bei Anfragen erlaubt, für welche die Einheiten nützliche Antworten sind, und die ein Wiederfinden verhindert, wenn die Einheiten für eine Anfrage nicht nützlich sind.[240] Dennoch müssen wir die Tatsache akzeptieren, dass es den perfekten Index nicht gibt. Zu viele Variablen sind bei einer Indexierung involviert, insbesondere die Individualität der Menschen. Gewisse Merkmale eines Index, die im Folgenden betrachtet werden, bürgen aber für eine akzeptierbare Indexqualität. Eigentlich könnten wir nun die Anforderungen an ein Schlagwortregister, die im Abschnitt 3.6 aufgelistet wurden, als Qualitätsmerkmale nochmals anführen. Da die einzelnen Anforderungspunkte sich wechselseitig beeinträchtigen und gegenseitige Zugeständnisse gemacht werden müssen, möchte ich die in der Fachliteratur aufgeführten Qualitätsmerkmale hier separat zusammentragen.

Ein Index besitzt eine Menge von charakteristischen Merkmalen, wenn das Wiederauffinden von Informationen so präzise und so vollständig wie möglich sein soll. Auch die im Folgenden skizzierten wichtigsten Merkmale können gegenseitig in Konflikt geraten.

1. Ein wichtiger Charakterzug eines „guten“ Index ist die Fähigkeit, das permanente Thema oder die Gegenstände eines Textes (*aboutness*, *topicality*) zu repräsentieren.[241] Vor allem in einem

[238] Fugmann 1999:202
[239] Cleveland/Cleveland 1990:143
[240] Lancaster 1998:77
[241] Moens 2000:70

allgemeinen Umfeld ist Gegenstandserkennung in Browsing-, Retrieval- oder Filtersystemen höchst wertvoll. Sie kann aber auch über den Kauf eines Buches entscheiden. Neben der Themenidentifizierung wird auch die Fähigkeit, die möglichen Bedeutungen eines Textes für seine Benutzer bzw. deren Textinterpretationen einzufangen, als Qualitätsmerkmal betrachtet. Dies kann mit einer detaillierteren Indexierung der Unterthemen oder bestimmter Informationen eines Textes realisiert werden.

2. Ein Index besteht oftmals in einer Reduzierung des Inhaltes des Originaltextes. Die Reduzierung ist das Resultat von Verallgemeinerung oder Selektion des Inhaltes. In Kontrast zu Punkt 1 ist das Merkmal der Reduktion von Bedeutung, wenn in grossen Dokumentkollektionen nach Informationen gesucht wird. Indexe, die eine Suche nach allgemeinen Konzepten möglich machen, besitzen ein zusätzliches Qualitätsmerkmal.

3. Das Ziel einer Indexierung ist nicht nur eine angemessene Beschreibung der Textinhalte, sondern die Differenzierung des Inhalts von den Inhalten anderer Texte. Dieser Charakterzug kollidiert mit Punkt 2.

4. Die Retrievalfähigkeit ist nur ein Aspekt von Indexqualität. Eine weitere Eigenschaft ist die Korrektheit. Fehler sind üblich, und es ist das Ziel eines Index, diese Fehler zu vermeiden, wobei die Unterlassungsfehler problematischer sind als die Vollzugsfehler (siehe Abschnitt 4.3).

5. Wirtschaftlichkeit gehört heute zu den Qualitätsmerkmalen von Indexen und ist ein unvermeidbares Kriterium, auch wenn daraus manchmal schlechte Indexe resultieren.[242] Dennoch ist der ausgefeilteste Index von höchster Qualität ohne Gebrauchswert, wenn ihn sich niemand leisten kann. Diese Eigenschaft steht mit allen anderen Punkten in Widerspruch.

Es liessen sich sicherlich noch weitere charakteristische Qualitätszüge finden, ich habe mich auf die mir am bedeutendsten erscheinenden konzentriert.

4.2 Faktoren, welche die Indexierungsqualität beeinflussen

Ein guter oder schlechter Index ist nicht das Resultat einer einzelnen Komponente, sondern von vielen Faktoren. Um zu zeigen, wie viele Faktoren das in der Tat sind, werden in Tabelle 3 die Faktoren aufgelistet, welche beispielsweise eine manuelle Indexierung beeinflussen können.

[242] Cleveland/Cleveland 1990:157

Indexiererfaktoren:	**Dokumentfaktoren:**	**Umfeldfaktoren:**
Fachgebietkenntnis Erfahrung Lese- und Verständnisfähigkeit Interesse Einstellung Konzentration persönliche Eigenheiten	Thema Komplexität Sprache und Ausdrucksweise Präsentation Zusammenfassung Länge	verfügbarer Computer Heizung/Klimaanlage Licht Lärm

Vokabularfaktoren:	**Prozessfaktoren:**
kontrolliert vs. frei Qualität des Vokabulars Normen Spezifität Terminologiekontrolle Begriffskontrolle lexikalische Mehrdeutigkeit Verfügbarkeit von Hilfsmitteln	Indextyp Indexierungsverfahren Regeln und Instruktionen verlangte Produktivität Indexierungstiefe Verwendungszweck Ziel der Indexierung

Tabelle 3: Faktoren, welche die Indexierungsqualität beeinflussen können[243]

Nach dem Herausarbeiten, was einen „guten“ Index charakterisiert und welche Komponenten Einfluss auf eine Indexierung ausüben, werden im folgenden Abschnitt die vier wichtigsten Kriterien zur Beurteilung eines produzierten Index erläutert.

4.3 Evaluation mit vier Bewertungskriterien

Generell kann ein Index als individuelle Einheit oder im Vergleich mit ähnlichen oder anderen Indexen bewertet werden.[244] Weiter kann er aufgrund seiner Übersichtlichkeit, seiner Nützlichkeit, seiner Eigenschaften usw. **direkt** bewertet werden oder **indirekt** anhand seiner Retrievalergebnisse. Diese beiden Ansätze der direkten und indirekten Evaluation sind die hauptsächlichen Vorgehensweisen beim Bewerten von Indexierungsergebnissen, wobei sich die erste Art der Qualitätsprüfung speziell für gedruckte oder kleinere, gut überschaubare Register und die zweite aufgrund der ausgesprochen grossen Indexdateien vor allem für kollektionsorientierte Indexierungen eignet. Wir wenden uns zuerst der ersten Art von Indexevaluierung zu, der direkten Bewertung von meist für den Druck bestimmten Indexen.

[243] Cleveland/Ceveland 1990:143; Lancaster 1998:83; Moens 2000:12
[244] Cleveland/Cleveland 1990:144

Indexe und andere Textrepräsentationen, die den Inhalt von Dokumenten für ein Retrieval erschliessen sollen, werden mit den Kriterien Exhaustivität, Spezifität, Korrektheit und Konsistenz beurteilt.[245]

Exhaustivität (*exhaustivity*) oder **Indexierungsbreite** ist das Ausmass der Abdeckung des fachlichen Inhalts des Dokumentes, der Grad, in welchem alle in einem Text enthaltenen Konzepte und Gegenstände erkannt und mit Indextermen beschrieben werden, inklusive der zentralen Themen sowie der nur kurz angesprochenen. Üblicherweise wird bei einem Vergleich mehrerer Indexe als Indikator für die Indexierungsbreite die durchschnittliche Anzahl der vergebenen Indexterme pro Dokument verwendet.

Spezifität (*specificity*) bezieht sich nicht nur auf den Grad der spezifischen Themenbehandlung, sondern auch auf den Grad der Verallgemeinerung in der Textrepräsentation. Das heisst, dass ein Index auch anhand der Möglichkeit von Anfragen nach allgemeinen Konzepten bewertet wird (*generic searches*). Gerade alphabetische Schlagwortregister leiden oft unter dem Vorwurf, dass sie die Suche nach Allgemeinbegriffen nicht fördern.

Indexierungsbreite und Indexierungsspezifität ergeben in Kombination die **Indexierungstiefe** als Bewertungskriterium. Diese lässt sich allerdings nur schwer operationalisieren und wird deshalb als Kombination der zwei unabhängigen Kriterien Exhaustivität und Spezifität, die sich leichter fassen lassen, betrachtet.[246] Eine hohe Indexierungstiefe liegt vor, wenn die vergebenen Indexterme die Themen eines Dokumentes sehr spezifisch und sehr exhaustiv treffen. Als Indikator für Indexierungstiefe wird üblicherweise die Dokumenthäufigkeit der Indexterme herangezogen (Anzahl der Dokumente in der Datenbasis, welche einen Indexterm enthalten).

Korrektheit (*correctness*) ist im Zusammenhang mit Indexen sehr wichtig. Indexieren ist anfällig für zwei Arten von Fehlern: **Unterlassungs-** oder **Auslassungsfehler** (*errors of omission*) und – ich nenne sie – **Vollzugs-** oder **Kommissionsfehler** (*errors of commission*).[247] Die erste Fehlerart bezieht sich auf Indexterme, die im Prozess des Indexierens zugeteilt hätten werden sollen, deren Zuordnung jedoch unterlassen blieb. Die zweite Fehlerart meint Indexterme, welche beim Indexieren eines Textes nicht zugeteilt hätten werden sollen, aber zugeteilt wurden. Beide Fehler geschehen bei der Inhaltsanalyse. Beim Übersetzen der erkannten Gegenstände in passende Indexterme können zwei weitere Fehler unterlaufen: 1. Es wird nicht der spezifischste der zur Verfügung stehenden Terme gewählt. 2. Ein nicht adäquater Indexterm wird aufgrund von fehlendem Fachwissen oder wegen Unsorgfältigkeit zugeteilt. Bei beiden Fehlschlägen handelt es sich um Spezialfälle von Fehlern, die gleichzeitig Unterlassungs- sowie Vollzugsfehler sind. Bei einer Qualitätsprüfung kann es relativ einfach sein, einen nicht korrekten Indexterm zu entdecken.

[245] Knorz 1997:136f.; Moens 2000:71
[246] Knorz 1997:137
[247] Lancaster 1998:79; Moens 2000:72

Unterlassungsfehler aber sind nicht immer offensichtlich, ausser der nicht zugeteilte Indexterm tritt beispielsweise im Titel auf.

Konsistenz (*consistency*) vergleicht Indexe, die von der gleichen Originalquelle in verschiedenen Kontexten, z.B. mit verschiedenen Techniken oder Verfahren, gemacht wurden.[248] Konsistenz kann auch für den Vergleich verschiedener Indexierer oder für den Vergleich von Indexen eines einzigen Indexierers verwendet werden. **Inter-Indexiererkonsistenz** (*inter-indexer consistency*) sagt aus, wie konsistent verschiedene Indexierer arbeiten, **Intra-Indexiererkonsistenz** (*intra-indexer consistency*) misst, wie konsistent ein Indexierer dasselbe Dokument zu verschiedenen Zeiten bearbeitet. Das Bewertungskriterium Konsistenz sollte nicht mit der Konsistenz, die beim Indexieren generell anzustreben ist, z.B. beim Zuteilen von möglichst spezifischen Termen oder beim Format und Layout von Registern, verwechselt werden.

Es ist offensichtlich, dass eine Qualitätsprüfung mit den vier diskutierten Kriterien ex post facto, also erst nach Vollendung der Indexierung vorgenommen wird. Gehen wir nun zu der indirekten Bewertung über, die eine Indexierungsevaluation mithilfe von Retrievaltestauswertungen vollzieht.

4.4 Evaluation mittels Retrievaltests

Da bei der retrievalorientierten Indexierung von riesigen Dokumentkollektionen eine manuelle Überprüfung der Indexe nicht möglich ist, werden die Qualitäten dieser Indexierungen durch Auswertungen von Retrievaltestergebnissen überprüft. Hierbei werden beispielsweise die Dokumente auf verschiedene Weise indexiert und ein vergleichender Retrievaltest durchgeführt. Eine Indexierung ist dann besser, wenn sich bessere Retrievalergebnisse herausstellen.

Meistens lässt man Information Retrieval-Systeme über Standardtestkollektionen laufen, um deren Performanz zu eruieren. Dabei gibt es eine vordefinierte Menge von Anfragen und die dazugehörigen relevanten Dokumente. Mit Recall und Precision werden die Ergebnisse dann ausgewertet (siehe Abschnitt 4.4.4). Je höher die Werte beider Masse, desto besser schneidet ein System ab. Der Testansatz der Berechnung der Performanz eines Informationssystems basiert auf der Annahme, dass, je effektiver ein System ist, desto besser kann es die Benutzerbedürfnisse befriedigen.[249]

Die Tests sind nur mit grossem empirischen Aufwand zu vollziehen und nach Ansicht von Knorz leider nicht zweifelsfrei anzuwenden.[250] Da beispielsweise ein Retrievalresultat bei kleinen Dokumentmengen nicht repräsentativ für grosse Dokumentbestände ist, sind aussagekräftige Tests

[248] Moens 2000:72
[249] van Rijsbergen 1979:120; Verdejo et al. 1999:9
[250] Knorz 1994:5

mit umfangreichen Stichproben von Fragen und Dokumenten mit riesigem Aufwand verbunden, sowohl in Organisation und Evaluierung. Viele methodische Probleme der Auswertung gelten heute noch nicht als befriedigend gelöst, und generell kann man kaum verbindlich sagen, wie man die Ergebnisse eines konkreten Tests verallgemeinern darf, z.B. auf andere Nutzerinteressen, andere Fachgebiete, andere Retrievaltechniken.[251]

Eine Indexierung ist genau dann besser als eine andere, wenn die damit erzielten Retrievalergebnisse besser sind. Sollen die Indexierungsergebnisse und die damit erreichte Retrievalqualität verbindlich bewertet werden, kann es keinen anderen Weg geben, als dass auch die Retrievaltests selbst einer Bewertung unterzogen werden. Nur so können verschiedene Indexierungsverfahren vergleichend bewertet werden. Die nachstehenden Punkte sollte man sich gemäss Knorz bei Retrievaltestevaluierungen stets vor Augen halten[252]:

- Rechercheergebnisse sind nicht nur vom gewählten Indexierungsverfahren abhängig, sondern auch von einer Vielzahl anderer Parameter: Sprache, Fachgebiet, Grösse und „Dichte“ der Datenbasis, Art der Retrievalfragen, verfügbare Retrievaloperationen, Ausgestaltung der Benutzerschnittstelle, Status und Kompetenz des Suchenden, Art der Relevanzbeurteilung.
- Ein Retrievaltest muss, wenn er verallgemeinerungsfähige Aussagen liefern soll, möglichst viele repräsentative Dokumente und Fragen einbeziehen. Dementsprechend aufwändig ist die Erzeugung von konkurrierenden Indexierungen und die Aus- und Bewertung der Retrievalergebnisse.
- Nicht nur die Konzeption eines Retrievaltests, sondern auch seine Auswertung wirft beachtliche methodische Probleme auf. Die Standardmasse der Bewertung von Retrievalantworten sind Precision und Recall (siehe Abschnitt 4.4.4). Einige diesbezüglich zu entscheidende Fragen sind: Wie sind diese Masse zu mitteln? Wie behandelt man Antworten ohne bekannte relevante Dokumente? Wie vergleicht man Antworten, bei denen ein Precision-Vorteil einem Recall-Vorteil gegenübersteht? Wie sichert man die gefundenen Qualitätsdifferenzen statistisch ab?

Allen Einwänden zum Trotz ist die Berechnung der Performanz eines Informationssystems ein wichtiger Faktor bei der Evaluierung von Informationswiedergewinnung. Viele Techniken und Masse wurden entwickelt und benutzt; jedes wurde entworfen, um einige Aspekte der Retrievalperformanz eines Systems zu bewerten. In den nun folgenden Abschnitten wird eine Auswahl dieser Methoden und Masse – im Hinblick auf die Auswertung der automatischen Indexierung der Buchausschnitte im zweiten Teil der Arbeit – beschrieben.

[251] Knorz 1994:5
[252] Knorz 1997:138

4.4.1 Fehlerstatistiken

Wenn eine Indexierungsmethode so beschaffen ist, dass ein „ideales“ Indexierungsergebnis verbindlich vorgegeben werden kann, kann die Abweichung von diesem vorgegebenen Standard in einer **Fehlerstatistik** erfasst werden.[253] Die fehlerhaften Indexierungsergebnisse werden manuell erfasst und klassifiziert. Eine zusammenfassende Beurteilung könnte dann beispielsweise aussagen, dass in vier Prozent aller Zuteilungsfälle Fehler auftraten und dass in 22% aller vorkommenden Fehler ein Fehler vom Typ „nicht spezifischster Term gewählt“ vorlag.

Durch eine Klassifikation der Fehler kann die Bewertung differenziert und eine Fehlerbehebung unterstützt werden. Allerdings bedingen manuell erstellte Fehlerstatistiken enorm viel Aufwand und sind für Indexierungen für grosse Datenbanken wenig geeignet.

Wenn ein Indexierungsstandard nicht mit letzter Verbindlichkeit vorgegeben werden kann, sollte man nicht von Fehlern, sondern nur von Abweichungen sprechen. Diese Abweichungen werden ausgezählt und beispielsweise zu einer Bewertung der Konsistenz mit dem vorgegebenen Standard umgerechnet.

4.4.2 Konsistenz

Um den oftmals gar nicht leistbaren Aufwand von umfassenden Retrievaltests zu umgehen, werden auch lokale und vorläufige Qualitätskriterien definiert, z.B. die Korrektheit des Verfahrens oder der Vergleich mit den Ergebnissen intellektueller Indexierung. Zu Letzterem wird zumeist der **Konsistenzfaktor q** oder **k** (*consistency factor*) verwendet, dessen Werte zwischen 0 (keine Gemeinsamkeit) und 1 (vollständig identisch) liegen und der wie folgt definiert ist:

$$q = \frac{|S \cap I|}{|S \cup I|}$$

Sei S die Menge der einzelnen Elemente im Indexierungsergebnis des Standards (Deskriptoren, syntaktische Relationen etc.), I die Menge der Elemente im zu bewertenden Indexierungsergebnis.[254] Zum Beispiel könnte S die Menge der automatischen Indextermzuteilungen und I die Menge der manuellen Indextermzuteilungen bezeichnen, die Konsistenz q berechnet sich zwischen S und I.

[253] Knorz 1997:136
[254] Knorz 1994:5f.; Knorz 1997:136f.

Das Konsistenzmass ist symmetrisch bezüglich S und I und kann deshalb auch für den Vergleich zweier beliebiger (vergleichbarer) Indexierungsergebnisse von gleichen Dokumenten verwendet werden. Sowohl Inter-Indexiererkonsistenz wie Intra-Indexiererkonsistenz können gemessen werden.

Konsistenzbewertungen sind vielfach als Bewertung automatischer Indexierungsverfahren mit manueller Indexierung als Standard verwendet worden. Ist die Konsistenz einer Indexierung mangelhaft, wird man kein gutes Retrieval erwarten und in jedem Falle auf Schwächen der Indexsprache oder der Bearbeitung schliessen können. Aber ob gute Konsistenz zu gutem Retrieval führt, hängt davon ab, wie diese Konsistenz erreicht wird. Es gibt vernünftige konsistenzsichernde Massnahmen, aber auch rein formale. Beispielsweise lässt sich die Regel „Indexiere alle Wörter mit weniger als zehn Buchstaben“ leicht befolgen, und sie ermöglicht optimale Konsistenz, ohne sinnvoll zu erfolgreichen Retrievalergebnissen beizutragen. Qualität und Konsistenz sind also nicht das Gleiche; eine Indexierung kann konsistent gut oder konsistent schlecht sein.[255] Nichtsdestotrotz gehören Konsistenz und Qualität intuitiv irgendwie zusammen. Da aussagekräftige Konsistenztests für die Beurteilung von Informationssystemen äusserst kostspielig und zeitraubend sind, existieren nur wenige verlässliche Testergebnisse. Für mehr Informationen zu der eher beschränkten Aussagekraft von Konsistenzwerten siehe Knorz 1997.[256]

Der Konsistenzfaktor wird im weiteren Verlauf der Arbeit keine Verwendung finden, im Gegensatz dazu jedoch die Relevanz, auf deren Eigenschaften und Problematik im nächsten Abschnitt eingegangen wird.

4.4.3 Relevanz

Das grundlegende Bewertungsmass der gefundenen Dokumente in einem Retrievaltest ist die **Relevanz** (*relevance*). Effektivität ist das Mass der Fähigkeit eines Systems, einen Benutzer im Hinblick auf die Relevanz der gefundenen Dokumente zu befriedigen.[257] Deshalb ist es zweckmässig, sich mit der Definition des Relevanzbegriffes näher auseinander zu setzen.

Nevelig/Wersig definieren Relevanz als „die Eigenschaft der bei Benutzung von Leistungen einer Informations- und Dokumentationseinrichtung in Frage kommenden Dokumentationseinheiten, um die der Benutzung zugrunde liegenden Benutzerbedürfnisse zu befriedigen“.[258] Die Relevanz bezieht sich also auf die Übereinstimmung zwischen einer bestimmten Anfrage und einem bestimmten Dokument aus der Sicht eines unabhängigen Jurors. Für Moens ist Relevanz ganz

[255] Lancaster 1998:85
[256] Knorz 1997:128, 137
[257] van Rijsbergen 1979:122
[258] so zitiert in Kaiser 1996:25

allgemein das Mass der Effektivität beim Kontakt zwischen einem Sender und einem Empfänger und entsprechend, etwas enger gefasst, das Effektivitätsmass bezüglich der Kommunikation im Information Retrieval. Die Relevanz ist dabei das Verhältnis eines Dokuments zu einem Benutzerbedürfnis, und diese spielt eine bedeutende Rolle bei der Befriedigung dieses Bedarfs.[259]

Spricht man von Relevanz bei Information Retrieval-Systemen, muss man zwischen der **Benutzerrelevanz** (*user relevance*) und der **Systemrelevanz** (*system relevance*) unterscheiden. Bei der Systemrelevanz ist die Entscheidung des Systems über den Nachweis eines Dokumentes bezüglich einer Query gemeint. Sie ist der Grad der formalen Übereinstimmung zwischen einer verschlüsselten Suchfrage und den Indextermen einer Dokumentationseinheit. Trifft der Benutzer diese Entscheidung, dann spricht man von der Benutzerrelevanz, die den Grad der vom Fragesteller angegebenen Übereinstimmung zwischen der Leistung der Informationseinrichtung und dem Benutzerbedürfnis angibt. Der Unterschied zwischen Benutzer- und Systemrelevanz kann als Indikator für den Nutzeffekt des Systems herangezogen werden.[260]

In Analogie zur Unterscheidung von Benutzer- und Systemrelevanz existieren für Cleveland/Cleveland zwei Ebenen von Relevanz.[261] Die erste Ebene meint die Relevanz per se. Die zweite Ebene wird als **Pertinenz** (*pertinence*) im Sinne von Sachdienlichkeit bezeichnet und basiert auf der Idee, dass bei der Evaluation von Sachindexen zwei Aspekte vorhanden sind: Erstens die Relevanz der in den Indextermen ausgedrückten wiedergewonnenen Informationen und zweitens die Nützlichkeit der Information für einen Benutzer, der eine Anfrage gestellt hatte. Relevanz designiert das Verhältnis zwischen einem Dokument und dem Index, Pertinenz das Verhältnis von Dokument und Benutzer. Der erste Aspekt von Relevanz kann quantifiziert werden, indem ein Fachexperte die Resultate evaluiert. Aber der einzige Weg, Pertinenz zu bewerten, ist die Befragung der Benutzer nach dem Gebrauchswert der gefundenen Informationen.

Die Benutzerrelevanz hat also etwas mit dem persönlichen Nutzen zu tun. Ein Dokument wird für einen Rechercheur dann relevant sein, wenn ihm das gefundene Dokument einen zusätzlichen Nutzen einbringt. Das heisst, ein und dasselbe Dokument kann für zwei Anwender, welche die gleiche Query gestellt haben, von unterschiedlichem Nutzen und damit auch von unterschiedlicher Relevanz sein in Abhängigkeit von der Vorbildung des Benutzers bzw. in Abhängigkeit, wie viele Dokumente bereits gefunden wurden. Sogar derselbe Benutzer kann in seiner Beurteilung von Relevanz zu verschiedenen Zeitpunkten Abweichungen aufweisen. Benutzerrelevanz kann sich mit der Zeit verändern und je nach Suchzweck variieren. Nur das, wonach von einem Fragesteller gesucht wird, ist relevant. Die Relevanz ist daher etwas Unscharfes und Subjektives. Damit werden aber auch die weit verbreiteten Masse Recall und Precision zu subjektiven Massen (siehe Abschnitt 4.4.4).[262] Auch wenn die Relevanz eine subjektive Grösse ist und verschiedene Benutzer sich in

[259] Moens 2000:13
[260] Kaiser 1996:25
[261] Cleveland/Cleveland 1990:147f.
[262] Kaiser 1996:25

ihrer Beurteilung von relevant oder nicht relevant in Bezug auf die gleichen Dokumente und die gleichen Anfragen grundlegend unterscheiden können, ist die Differenz dennoch nicht gross genug, um Experimente mit Dokumentkollektionen, für welche Textanfragen mit korrespondierenden Relevanzschätzungen zur Verfügung standen, wertlos zu machen.[263]

Relevanz spielt in jedem Retrievalsystem eine wichtige Rolle. Jede Anfrage an ein Informationssystem partitioniert eine Dokumentkollektion in relevante und nicht relevante Dokumente sowie in gefundene und nicht gefundene Dokumente.[264]

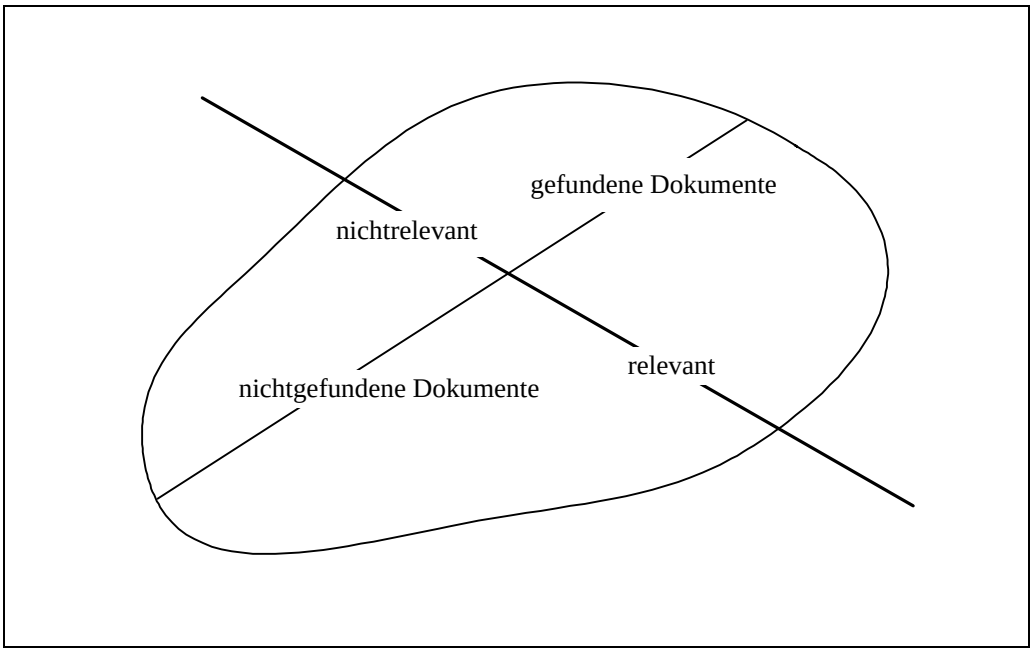

Abbildung 2: Die Aufteilung aller Dokumente bei einer Anfrage

Als Kriterium für Relevanz eines Retrievalergebnisses gilt gemäss van Rijsbergen: Ein Dokument ist relevant für ein Informationsbedürfnis, wenn es mindestens einen Satz beinhaltet, der für dieses Bedürfnis relevant ist.[265] Diese Beurteilung von Relevanz, wonach ein Dokument das Informationsbedürfnis des Benutzers erfüllt, nennt Schneider **subjektive Beurteilung**. Bei einer **objektiven Beurteilung** wird ein Dokument von einem Experten oder einem Komitee als relevant bezüglich einer bestimmten Anfrage definiert.[266] Die subjektive Beurteilung bezieht sich auf die Benutzerrelevanz oder Pertinenz, die objektive Beurteilung auf die Systemrelevanz. Wir wollen uns in Anlehnung an Schneider darauf festlegen, dass für uns von jetzt an Relevanz „das Mass der

[263] van Rijsbergen 1979:121
[264] Salton 1989:249
[265] van Rijsbergen 1979:121
[266] Schneider 2001:4

Übereinstimmung zwischen einem Dokument und der Suchanfrage aus der Sicht eines Experten“ ist.[267]

Für die Bewertung der Termgewichtungsmethode mithilfe der abgeänderten inversen Dokumentfrequenz in Kapitel 5 ist die Diskussion um die verschiedenen Relevanzauffassungen von geringer Bedeutung, da die Menge der relevanten Objekte aufgrund der vorhandenen idealen Indexe klar bestimmbar ist, auch wenn diese Buchindexe natürlich nicht generell als das Mass aller Dinge angesehen werden dürfen.

Auf der in unserem Sinne definierten Relevanz basieren alle diejenigen quantitativen Masse zur Evaluation von Retrievaltestergebissen, welche in den nachstehenden Abschnitten vorgestellt werden. Die am häufigsten verwendeten Masse zur Messung der Retrievaleffektivität oder Retrievalperformanz sind Recall und Precision.

4.4.4 Recall und Precision

Eine Vorbemerkung: Da die beiden Begriffe *Recall* und *Precision* schon sehr fortgeschritten in den deutschen Wortschatz übergegangen sind, werde ich die englischen Begriffe anstelle der deutschen Termini *Ausbeute* und *Präzision* verwenden.

Die Effektivität eines Retrievalsystems wird mit Precision und Recall gemessen. Mit anderen Worten: Die beiden Masse messen die Fähigkeit eines Systems, relevante Dokumente hinsichtlich einer Anfrage zu finden und gleichzeitig die nicht relevanten zurückzuhalten. Dahinter verbirgt sich die Annahme, je effektiver ein System, desto besser die Zufriedenstellung der Benutzerbedürfnisse. Nach Auffassung von van Rijsbergen sind Recall und Precision die Masse, die sich zur Messung der Effektivität gut eignen.[268] Sie basieren auf einer binären Beurteilung von Relevanz.

Der **Recall** stellt das Mass für die Vollständigkeit des Retrievalergebnisses dar und ist definiert als das Verhältnis zwischen den gefundenen relevanten Dokumenten und der Gesamtzahl der im Dokumentenbestand vorhandenen relevanten Dokumente.

$$R = \frac{\text{Anzahl der gefundenen relevanten Dokumente}}{\text{Gesamtanzahl der relevanten Dokumente}}$$

[267] Schneider 2001:4
[268] van Rijsbergen 1979:120

Der Wertebereich des Recalls geht von 0 bis 1. Ein Recall von 0 wird für das schlechteste Ergebnis, 1 für das bestmögliche vergeben.

Erinnern wir uns an die Zerlegung der Dokumentkollektion bei einer Anfrage und stellen wir diese etwas anders dar, so ergeben sich für die Dokumente in einer Kollektion die folgenden Möglichkeiten[269]:

	relevant	nicht relevant	
gefunden	a	b	a + b
nicht gefunden	c	d	c + d
	a + c	b + d	e

Tabelle 4: Quantitäten der Dokumente in einer Kollektion

Aus dieser Tabelle und der allgemein formulierten Formel für Recall von oben ergibt sich die genaue Recall-Formel:

$$R = \frac{a}{a + c}$$

Die Gesamtanzahl aller relevanten Dokumente in einem Dokumentenbestand wird dargestellt als die Anzahl der gefundenen relevanten Dokumente vermehrt um die Anzahl der nicht gefundenen relevanten Dokumente. a (die Grösse der gefundenen relevanten Dokumente) ist leicht bestimmbar. Für die Grösse c (die Grösse der nicht gefundenen relevanten Dokumente) muss in einem unüberschaubar grossen Informationssystem ein Schätzwert angenommen werden, um den Nenner des Recalls bestimmen zu können. Für diesen Schätzvorgang wurden verschiedene Methoden entwickelt, um eine möglichst genaue Annäherung an die Gesamtzahl aller relevanten Dokumente bzw. der im System verbleibenden Dokumente c zu erhalten. In den meisten Fällen wurde ein relativ kleines Subset von Anfragen und eine Untermenge des Dokumentenbestandes vollständig auf relevantes Material hin durchsucht und diese Zahl dann auf den tatsächlichen Dokumentenbestand hochgerechnet.[270] Ich werde hier nicht genauer darauf eingehen, da die Anzahl der nicht gefundenen relevanten Indexterme aus den Schlagwortverzeichnissen der untersuchten Bücher für die Auswertungen im Teil II klar bestimmbar ist.

[269] van Rijsbergen 1979:122
[270] Kaiser 1996:26

Was sagt der Recall als Vollständigkeitsmass aus? Nehmen wir an, wir haben in einer Kollektion total 100 Dokumente. Bezüglich einer bestimmten Anfrage sind 75 Dokumente relevant und 25 nicht. Auf genau diese Anfrage erhält ein Benutzer als Rechercheresultat 50 Dokumente. Dann beträgt der Recall oder die Ausbeute für dieses Ergebnis 50 dividiert durch 75, also aufgerundet 0,667. Diese Zahl bedeutet, dass vom System 66,7% aller relevanten Dokumente in Bezug auf die Anfrage gefunden wurden.

Das zweite wichtige Mass zur Messung der Retrievaleffektivität ist die **Precision** oder Präzision. Die Precision dient zum Messen der Genauigkeit der Suche und als Indikator für die Fähigkeit eines Information Retrieval-Systems, nicht relevante Dokumente auszuscheiden. Die Precision ist definiert als das Verhältnis der gefundenen relevanten Dokumente zur Anzahl aller gefundenen Dokumente.

$$P = \frac{\text{Anzahl der gefundenen relevanten Objekte}}{\text{Gesamtanzahl der gefundenen Objekte}}$$

Auch der Wertebereich der Precision geht von 0 bis 1. Je besser die Trefferquote, desto näher ist der Wert bei 1. Das Ziel ist also, den Precisionwert zu maximieren (das Gleiche gilt für den Recall).

In Bezug auf die Dokumentquantitäten der Tabelle 4 von oben ergibt sich die folgende Formel für die Precision:

$$P = \frac{a}{a + b}$$

Wenn wir in einer Kollektion ein Total von 100 Dokumenten haben und bezüglich einer bestimmten Anfrage 25 relevante Dokumente gefunden werden plus zusätzlich 15 nicht relevante, dann beträgt die Precision für dieses Ergebnis 25 dividiert durch 40, also 0,625. Diese Zahl bedeutet, dass vom System 62,5% aller gefundenen Dokumente in Bezug auf eine Anfrage relevant sind.

Nur eine Betrachtung beider Masse, Recall und Precision, ist sinnvoll,[271] da der Recall die Ballastquote unberücksichtigt lässt und leicht auf das Maximum 1 gesetzt werden kann, indem alle im Dokumentenbestand vorhandenen Dokumente nachgewiesen werden. In diesem Fall wäre dann allerdings der Wert der Precision sehr niedrig. Eine alleinige Betrachtung der Precision würde

[271] Kaiser 1996:27

nichts über die Vollständigkeit des Retrievalergebnisses aussagen. Die Precision könnte dadurch maximiert werden, dass sehr wenige Dokumente nachgewiesen werden.

Daher wird bei vielen Evaluierungen von Retrievalexperimenten der **Recall-Precision-Graph** (*recall-precision-curve*) verwendet. In diesem Graph wird auf der y-Achse die Precision und auf der x-Achse der Recall eingetragen und so versucht, ein Bewertungsmass zu schaffen, dass beide Grössen mit einbezieht. Abbildung 3 zeigt den typischen Verlauf solch eines Graphen.

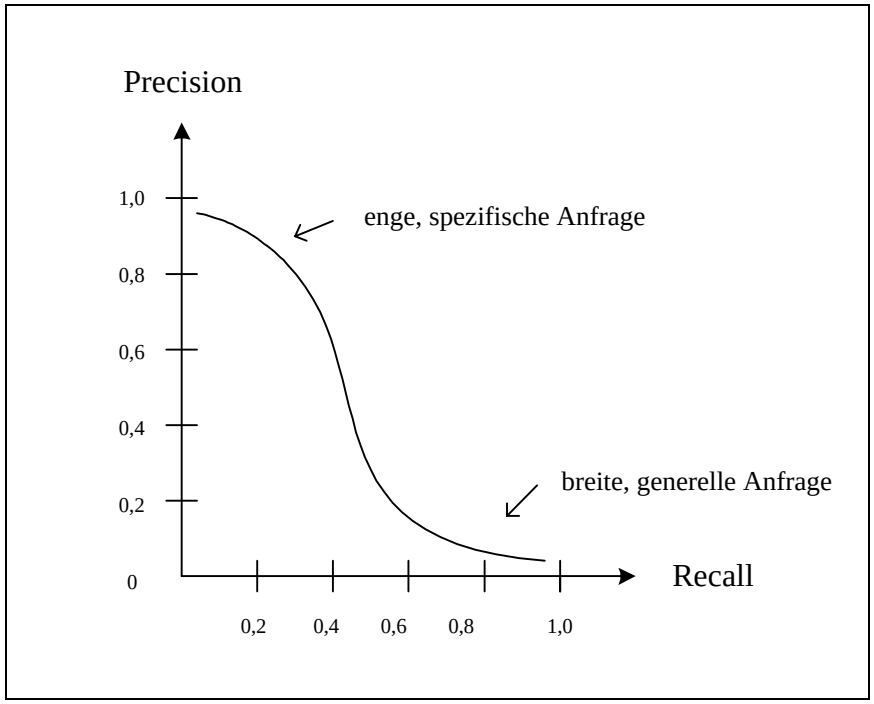

Abbildung 3: Typische Precision-Recall-Kurve[272]

Die Performanz einer jeden Anfrage wird mit einer derartigen Kurve dargestellt. Um die allgemeine Performanz eines ganzen Systems zu messen, wird die Menge aller Kurven kombiniert und eine Durchschnittskurve produziert.[273] Für die Berechnung von Durchschnittkurven gibt es diverse Techniken. Ich werde hier nicht näher darauf eingehen.[274]

Recall und Precision werden sowohl von Seiten der Indexierung als auch von Seiten des Retrievals, also von der Formulierung der Suchanfrage beeinflusst. Die Suchstrategie des Benutzers und die Formulierung der Query wirkt sich auf beide Effektivitätsmasse aus. Für den Bereich der Indexierung kann gesagt werden, dass Exhaustivität und Spezifität zwei Faktoren sind, die das

[272] Frei 1999:3
[273] van Rijsbergen 1979:123
[274] Für weitere Informationen dazu siehe van Rijsbergen 1979:123-149

Resultat einer Recherche beeinflussen. Eine sehr erschöpfende Indexierung, die eine grosse Anzahl Indexterme vergibt, ist einerseits mit einer Erhöhung des Recalls verbunden, andererseits jedoch mit einer Reduzierung der Precision. Umgekehrt geht hohe Spezifität der Indexierung, also die Mächtigkeit, Themen und Gegenstände sehr präzise zu beschreiben, mit einer Erhöhung der Precision zu Ungunsten des Recallwertes einher. Das heisst, dass Recall und Precision aus der Sicht der Indexerstellung prinzipiell in einem inversen Verhältnis zueinander stehen und dass ein optimaler Grad an Indexierungsexhaustivität und -spezifität angestrebt wird.[275] Das letztendliche Ziel einer Indexierung ist eine Erhöhung sowohl des Recalls als auch der Precision sowie eine Ausbalancierung der zwei Werte. In den meisten Fällen wird ein maximaler Recall zu erreichen versucht bei einem dennoch akzeptablem Grad an Precision.[276]

Die Vorteile der Standardmasse Recall und Precision sind ihre weite Verbreitung und die einfache Interpretierbarkeit.[277] Aus diesen Gründen werden die beiden Masse auch für die Auswertung der Daten im statistischen Teil der Arbeit herangezogen.

4.4.5 Fallout

Neben Recall und Precision gibt es das weitaus seltener verwendete Mass **Fallout**.[278] Der Fallout stellt das Mass für den Ballast eines Retrievalergebnisses dar und ist definiert als das Verhältnis der gefundenen nicht relevanten Dokumente zur Gesamtzahl aller nicht relevanten Dokumente im Dokumentenbestand.

$$F = \frac{b}{b + d}$$

Der Fallout misst die Fähigkeit des Systems, mit nicht relevanten Dokumenten umzugehen, und sollte möglichst nahe bei 0 sein (das Maximum des Wertes kann 1 betragen). Die Aussagekraft ist für einen Benutzer jedoch eher gering, da der Fallout primär systembezogen zu verstehen ist. Sind Recall und Precision bereits ermittelt, ist natürlich auch der Fallout leicht zu bestimmen, da die Anzahl der nicht gefundenen relevanten Dokumente bekannt ist. Im weiteren Verlauf wird neben Recall und Precision auch der Fallout von Bedeutung sein, da die Methode der inversen Seitenfrequenz auch daraufhin getestet werden soll, welche nicht relevanten Indexterme erwischt werden.

[275] Cleveland/Cleveland 1990:150; Kaiser 1996:27
[276] Lancaster 1998:78; Moens 2000:70; Salton 1989:277ff.
[277] Kaiser 1996:27; van Rijsbergen 1979:120
[278] Kaiser 1996:2; van Rijsbergen 1979:122

Wir wollen das Mass des Fallout wiederum an einem Beispiel demonstrieren. Bei einem Suchauftrag werden 25 nicht relevante Dokumente geliefert (b) und 40 nicht relevante Dokumente korrekterweise zurückbehalten (d). Das ergibt ein Total von 65 nicht relevanten Dokumenten hinsichtlich der gestellten Anfrage (b + d). Der Fallout beträgt aufgerundet 0,385 (25 dividiert durch 65). Somit werden 38,5% aller nicht relevanten Dokumente als Ballast an den Benutzer übergeben.

4.4.6 Accuracy

Das Mass der **Accuracy** sagt etwas über die Treffgenauigkeit der Antworten bzw. der gefundenen Dokumente aus. Accuracy wird bei der Performanzmessung von Systemen vor allem im Umfeld der Textkategorisierung verwendet, im Information Retrieval eher selten.[279] Die Anzahl aller gefundenen relevanten oder korrekterweise nicht gefundenen nicht relevanten Dokumente wird in ein Verhältnis zu der Gesamtanzahl aller Dokumente in einer Kollektion gesetzt. Der Wertebereich liegt zwischen 0 und 1. Je höher der Wert, desto besser die Retrievaleffektivität in Bezug auf die Treffgenauigkeit.

$$A = 1 - \text{Fehlerrate} = \frac{a + d}{a + b + c + d}$$

Ein Accuracywert von 0,627 bedeutet, dass bei einer Suchanfrage 62,7% Treffgenauigkeit erreicht werden, dass 62,7% der relevanten Dokumente gefunden und der nicht relevanten Dokumente korrekterweise nicht gefunden werden.

4.4.7 Error

Error oder **Fehlerrate** ist ein Mass, das der Messung der Fehlschläge dient. Es bezieht sich sowohl auf Unterlassungs- wie auch auf Vollzugsfehler, und sein Wert sollte möglichst nahe bei 0 sein.

$$E = 1 - \text{Accuracy} = \frac{b + c}{a + b + c + d}$$

[279] Aas/Eikvil 1999:22; Moens 2000:105

Die Fehlerrate definiert sich als das Verhältnis aller gefundenen nicht relevanten plus aller nicht gefundenen relevanten Dokumente zur Anzahl sämtlicher Dokumente in einer Kollektion. Eine Fehlerrate von 0,373 heisst, dass 37,3% aller Dokumente bei einer Recherche fälschlicherweise geliefert oder zurückbehalten wurden. Wie wir aus den Formeln ersehen können, ergeben Accuracy und Error zusammen 1 bzw. 100%.

4.4.8 E-Wert

Um ein einziges Mass der Effektivität zu erhalten, wurde der **E-Wert** entwickelt. Er kombiniert Recall und Precision und sollte einen Wert möglichst nahe bei 0 erzielen.[280]

$$\text{E-Wert} = 1 - \frac{(\beta^2 + 1) * P * R}{\beta^2 * P + R}$$

P bezieht sich auf die Precision, R auf den Recall, und β ist ein Faktor, der die relative Wichtigkeit von Recall und Precision angibt.

4.4.9 F-Wert

Damit bei einem Mass zur Messung der Retrievaleffektivität Recall und Precision von gleicher Wichtigkeit sind (β = 1) und höhere Werte (idealerweise 1) mit einer besseren Effektivität korrespondieren, wurde auf der Grundlage des E-Wertes der **F-Wert** definiert.[281] Er kombiniert wie der E-Wert die Masse Recall und Precision

$$F_{\beta = 1} = 1 - \text{E-Wert}$$

Zu E-Wert und F-Wert werde ich keine zusätzlichen Erläuterungen anfügen, denn sie werden keine weitere Verwendung finden. Recall, Precision, Fallout, Accuracy und Error werden wir im statistischen Teil wieder antreffen, obschon sämtliche Masse zur Berechnung der Effektivität nach

[280] Moens 2000:105; Schneider 2001:3f.
[281] Moens 2000:105; Schneider 2001:3f.

Ansicht von van Rijsbergen eine bessere Determinierung von Relevanz benötigen würden.[282] In unserem Fall, wo die realen Schlagwortverzeichnisse als Idealindexe betrachtet werden, kommt die Problematik einer adäquaten Relevanzbestimmung nicht zum Tragen. Als relevante Indexterme gelten diejenigen, die im realen Index stehen. Die Masse zur Effektivitätsberechnung können somit ohne schwerwiegende Bedenken eingesetzt werden.

Damit habe ich den ersten, theoretischen Teil der Arbeit, welcher einen möglichst umfassenden Überblick über verschiedene Indexe, das Indexieren von Dokumenten oder Büchern, die üblichen Indexierungsverfahren sowie die Methoden zur Evaluierung von Indexen geben sollte, abgeschlossen. Wir begeben uns nun zum zweiten und statistischen Teil.

[282] van Rijsbergen 1979:120

TEIL II

5 Statistische Untersuchungen

In diesem statistischen Teil der Arbeit werden fünf verschiedene Buchindexe einer detaillierten Untersuchung unterzogen, um zu ermitteln, „was“ in einem realen Buchregister steht. Dazu werden zunächst die Anzahl der Schlagwörter, Untereinträge und Tokens, die in den Verzeichnissen vorkommen, festgehalten, und es wird eruiert, aus welchen lexikalischen Wortkategorien und welchen syntaktischen Bestandteilen die einzelnen Einträge der fünf Register bestehen. Ferner wird im dritten Unterkapitel erforscht, wie gut sich die inverse Seitenfrequenz, eine abgewandelte Form der inversen Dokumentfrequenz, zur automatischen Indexierung von deutschsprachigen Texten eignet, indem die Nominal- und Präpositionalphrasen aus fünfzigseitigen Ausschnitten der Textkörper von den fünf untersuchten Büchern extrahiert, mit der inversen Seitenhäufigkeit gewichtet und die damit ermittelten Indextermkandidaten mit den Einträgen in den dazugehörigen Indexen verglichen werden. Es werden also Schlagwörter automatisch extrahiert und zugeteilt, eine Art Schlagwortretrieval betrieben.

5.1 Material und Methoden

Die untersuchten Bücher sind alle in deutscher Sprache verfasst und stammen aus den Bereichen Linguistik, Computerlinguistik, Informatik und Philosophie:[283]

- Cap, Clemens. H.:
  Theoretische Grundlagen der Informatik.

- Hausser, Roland:
  Grundlagen der Computerlinguistik. Mensch-Maschine-Kommunikation in natürlicher Sprache.

- Linke, Angelika/Nussbaumer, Markus/Portmann, Paul R.:
  Studienbuch Linguistik.

- Stegmüller, Wolfgang:
  Hauptströmungen der Gegenwartsphilosophie. Eine kritische Einführung.

[283] vgl. Literaturverzeichnis: Verzeichnis der untersuchten Bücher

- Werlen, Iwar:
  Sprache, Mensch und Welt. Geschichte und Bedeutung des Prinzips der sprachlichen Relativität.

Alle Bücher beinhalten *expository text*, wo der Akzent auf den Themen und Unterthemen eines Textes liegt. Im Gegensatz dazu fokussieren *narrative texts* den Plot einer Geschichte.[284] Die Sachbereiche wurden nicht zufällig, sondern ganz bewusst ausgewählt im Hinblick auf das letztendliche Ziel dieser Arbeit, Vorlesungsskripte der Computerlinguistik zu indexieren. Da die Computerlinguistik sowohl Linguistik, Informatik und auch Philosophie (insbesondere die Logik) vereint, dürfen in einem Skript der Computerlinguistik ähnliche Gegenstände, Themen sowie auch Fachausdrücke, Formeln, Symbole etc. erwartet werden. Ein weiteres Kriterium zur Auswahl der Bücher war die Länge sowie die Nützlichkeit der Schlagwortverzeichnisse. Hinsichtlich des Gebrauchswertes wurden Empfehlungen von verschiedenen Personen der Abteilung Computerlinguistik an der Universität Zürich hinzugezogen. Die Längen der einzelnen Schlagwort- und Namenverzeichnisse sind in der Tabelle 5 dargestellt.

	Total der Buchseiten	Anzahl der Seiten des Namenindex	Anzahl der Seiten des Sachindex	Total der Seiten des Gesamtindex	Anteil des Gesamtindex in % zum Buchseitentotal
Cap	332	-	14	14	4,22%
Hausser	572	4	16	20	3,50%
Linke et al.	463	-	7	7	1,51%
Stegmüller	548	3	10	13	2,37%
Werlen	273	5	24	29	10,62%

Tabelle 5: Länge der untersuchten Buchindexe

Im Abschnitt 3.5 „Form und Layout eines Schlagwortregisters“ haben wir gesehen, dass die Anzahl der Registerseiten bei Büchern allgemeiner Fachliteratur ca. 7-8% der Buchgesamtseitenzahl betragen sollte. Die Gesamtregister der untersuchten Bücher (Namen- und Sachregister zusammen) erreichen diese Vorgabe jedoch nur in einem Fall. Es hat sich bei der Selektion des Materials ergeben, dass die Indexe von zahlreichen Sachbüchern auf nur etwa 2 bis 3% Seitenanteil kommen. Insofern sind die Umfänge von drei der gewählten Indexe von eher überdurchschnittlicher Länge. Dieser Umstand lässt darauf schliessen, dass auf die Erstellung der Register relativ grossen Wert gelegt wurde und dass sich die Bücher deshalb für die präsentierte Untersuchung eignen. Die Bücher von Linke et al. und Stegmüller wurden in die Untersuchung mit einbezogen, da sich deren Register als besonders übersichtlich und nützlich herausgestellt haben.

[284] Moens 2000:29

Das Postulat von 6 bis 8 Registereinträgen pro Buchseite gemäss Mulvany (siehe Abschnitt 3.5) wird von den für die Untersuchung selektierten Büchern in keinem einzigen Fall erfüllt. Tabelle 6 zeigt die Anzahl der Registereinträge bzw. Schlagwörter pro Buchseite bei den fünf Autoren.

	Anzahl Buchseiten des Textkörpers	Anzahl Namen-registereinträge pro Buchseite	Anzahl Sach-registereinträge pro Buchseite	Anzahl Gesamt-registereinträge pro Buchseite
Cap	303	-	2,46	2,46
Hausser	534	0,45	1,78	2,23
Linke et al.	435	-	1,60	1,60
Stegmüller	518	0,29	0,94	1,23
Werlen	213	1,27	3,89	5,16

Tabelle 6: Anzahl Registereinträge pro Buchseite

Selbst das gesamte Register von Werlen kann die Forderung nicht erfüllen, obschon dieses Register ausgesprochen lang ist und 10,62% des Buchseitentotals ausmacht. Ich schliesse daraus, dass 6 bis 8 Schlagwörter pro Buchseite eindeutig zu viel sind und dass die Länge eines das Postulat erfüllenden Registers bezüglich dessen Nützlichkeit fragwürdig erscheint. Der Durchschnitt der erforschten Bücher ergibt den angemesseneren Wert von 2,54 Schlagwörter pro Buchseite.

Ursprünglich war es die Absicht, auch ein Buch aus der Fachwelt des Indexierens auszuwählen, da gerade diese Bücher doch über einen entsprechend qualitativ hochstehenden Index verfügen sollten. Die Idee wurde aufgegeben, da die meisten dieser Bücher in Englisch verfasst sind und da das Sachverzeichnis des in deutscher Sprache geschriebenen Buches von Fugmann zum Thema „Inhaltserschliessung durch Indexierung“ zwar 7% der Buchseiten ausmacht, sich jedoch als nicht mit den anderen Registern vergleichbar herausstellte, weil das Verzeichnis aus verhältnismässig wenigen Schlagwörtern, jedoch ausserordentlich vielen dazugehörigen Untereinträgen besteht, was das Register eher zu einem systematischen Register macht; die Indexe der anderen untersuchten Bücher jedoch eindeutig alphabetische Register darstellen.[285] Ferner wurde das Sachverzeichnis von Fugmann als unübersichtlich und umständlich in Bezug auf Recherchen beurteilt, was ebenso auf die Tatsache zurückzuführen ist, dass das Register nur wenige Schlagwörter oder Zugangspunkte, wonach schnell gesucht werden könnte, umfasst.

Für die Untersuchung der Schlagwortregister wurden die fünf Gesamtverzeichnisse der ausgewählten Bücher von Hand durchgekämmt und die Zuteilung der Wortkategorieklassen sowie

[285] Fugmann 1999:224-243

die Erfassung der syntaktischen Bestandteile der Einträge entsprechend manuell durchgeführt. Zur Überprüfung des Gebrauchswertes der inversen Seitenfrequenz wurden je 50 Textkörperseiten der fünf Bücher eingescannt und mit einer Texterkennungssoftware überarbeitet sowie manuell kontrolliert. Mit dem NP-Chunker von Wojciech sind die Nominal- und die Präpositionalphrasen maschinell aus den Texten sowie aus den dazugehörigen Tabellen und Grafiken extrahiert worden, wobei auch die untergeordneten Nominal- oder Präpositonalphrasen aus den komplexen Phrasen ermittelt wurden.[286] Überschriften, Topicsätze und Text wurden nicht gesondert behandelt. Auch der Verarbeitungsschritt der Phrasenerkennung wurde von Hand nachgeprüft, um die Verlässlichkeit der Daten und Auswertungen zu gewährleisten. Die extrahierten Nominal- und Präpositionalphrasen wurden danach mit einer selbst angefertigten Stoppwortliste ausgefiltert, lemmatisiert und normalisiert (z.B. wurden alle Adjektivattribute durch Komma abgetrennt nach dem Kopfwort einer Phrase aufgeführt) sowie alphabetisch sortiert, sodass sämtliche identischen oder ähnlichen Indextermkandidaten identifiziert und das Gewicht der inversen Seitenfrequenz für jeden Indextermkandidaten errechnet werden konnte. Wiederum habe ich alle erzielten Werte der inversen Seitenfrequenz manuell nachkontrolliert.

Zum besseren Verständnis, wie das Datenmaterial verarbeitet wurde, fasse ich die einzelnen Verarbeitungsschritte nochmals zusammen.

Verarbeitung der Buchregister (massgebend für Abschnitt 5.2):

1. Einscannen der fünf Buchregister
2. Texterkennung mittels Texterkennungssoftware
3. Auszählung der Schlagwörter und der Untereinträge
4. Auszählung der Tokens
5. Bestimmung der lexikalischen Wortkategorien der Schlagwörter und Untereinträge
6. Bestimmung der lexikalischen und syntaktischen Bestandteile der Schlagwörter und der Untereinträge

Verarbeitung der Textkörperausschnitte (massgebend für Abschnitt 5.3):

1. Einscannen der fünfzigseiten Textausschnitte
2. Texterkennung mittels Texterkennungssoftware
3. Nominal- und Präpositionalphrasenextraktion mit dem NP-Chunker von Wojciech
4. Ausfilterung der Funktionswörter mittels Stoppwortliste
5. Lemmatisierung der inhaltstragenden Wortformen
6. Umstellung der Wortreihenfolge auf eine kanonische, invertierte Form
7. alphabetische Sortierung der Wortgruppen
8. Berechnung der inversen Seitenhäufigkeit für jede nominale Wortgruppe

[286] Wojciech 1999

Sämtliche Verarbeitungsschritte sowohl für die Analyse der Register wie auch für die Untersuchung der Textausschnitte wurden manuell nachgeprüft.

5.2 Untersuchung der Schlagwortregister

Dieser Abschnitt verfolgt das Ziel, herauszuarbeiten, was tatsächlich in realen Buchregistern steht. Es wird ergründet, ob sich insbesondere Substantive als Schlagwörter eignen, wie das von Mulvany vorgeschlagen wird (vergleiche Abschnitt 3.2.1.1), und wie komplex die einzelnen Einträge aufgebaut sind. Zuerst werden die Mengenverhältnisse von Schlagwörtern, Untereinträgen sowie der Tokens studiert.

5.2.1 Die Schlagwörter und Untereinträge

In den untersuchten Registern finden sich Schlagwörter und Untereinträge. In der folgenden Tabelle 7 werden die Mengen der Schlagwörter und Untereinträge aufgelistet. Hierbei entspricht die Anzahl der Schlagwörter der Anzahl der Einträge eines Registers.

Anzahl	Cap	Hausser	Linke at al.	Stegmüller	Werlen	Total
Schlagwörter Namen-register	-	242	-	149	270	661
Untereinträge Namen-register	-	0	-	0	0	0
Schlagwörter Sachregister	744	948	693	487	829	3701
Untereinträge Sachregister	361	238	0	205	630	1434
Schlagwörter Gesamt-register	744	1190	693	636	1099	4362
Untereinträge Gesamt-register	361	238	0	205	630	1434

Tabelle 7: Anzahl Schlagwörter und Untereinträge in den untersuchten Registern

Die Zusammenführung von Namenregister und Sachregister zu einem Gesamtregister bei den drei Büchern, die beide Registertypen besitzen, wird vorgenommen, weil in den zwei Registern, die kein separates Namenverzeichnis aufweisen, ebenso Personennamen wie Schlagwörter enthalten sind. Die fünf verschiedenen Register lassen sich durch eine Gesamtbetrachtung somit besser vergleichen.

Die Tabelle zeigt, dass in sämtlichen Namenregistern keine Untereinträge vorkommen. Bei Linke et al. finden sich auch im Sachverzeichnis keine Untereinträge. Hier werden alle Ordnungswörter stets wiederholt, auch wenn sich ein Ordnungswort als Schlagwort und die dazugehörigen Angaben als Untereinträge mit dem Verzicht auf die Wiederholung des Schlagwortes darstellen lassen würden. Es kann also nach jeder einzelnen Angabe gesucht werden, und somit sind alle Einträge auch Schlagwörter (nach unserer Definition in Abschnitt 3.2.1.1 sind es die Schlagwörter, welche alphabetisch sortiert werden und die Zugangsstellen bilden). Diese Art der Darstellung der Einträge bei Linke et al. ist übrigens meiner Auffassung nach für einen Benutzer sehr übersichtlich und ermöglicht eine bequeme Suche. Die folgenden Auszüge aus dem Register von Linke et al. und dem von Werlen, bei welchen die Darstellung den Originalen entspricht, dient der Veranschaulichung:

Linke at al.:

Abtönungspartikel, 275
Abwandlung, innere, 71
Akzent, 423
adjacency pair, 280
Adjunktion, 120
Affix, 61
Akt, initiierender, 279
Akt, respondierender, 279
Aktant, 81, 112
Aktualgenese, 330
Allomorph, 68
Allophon, 68, 385, 390, 426
Alltagsbegriff, 157, 345, 352
Alter, 312
Ambiguität, 141
Ambiguität, syntaktische, 141
Amplitude, 414
Analphabetismus, 393

Werlen:

Ableitung, deverbale 121
Abstrakta 119
Abstraktion 84
Adjektiv 119
- (als erstes Wort) 37 (A 29)
-, attributives 81
-, prädikatives 80
Adjektivstellung 81 (A 75)
Adverb 80
Äquivokation 25
Ästhetik 103. 104
äussere (Laut-)Form 60 s. Form, Laut-Form
Agglutination 108
Akademie der Wissenschaften, Berlin 24. 31. 34
Akkusativ 121. 122
-, inhumaner 120-122

Im Gesamtregister von Werlen steht bei jedem 1,74. Schlagwort ein Untereintrag. Bei Cap tritt jedes 2,06. Schlagwort ein Untereintrag auf, bei Stegmüller jedes 3,10. Schlagwort und bei Hausser genau jedes 5. Schlagwort. Hausser verwendet somit am wenigsten Untereinträge in Relation zu der Anzahl Schlagwörter (ausser natürlich Linke et al.). Umgekehrt ausgedrückt stehen im Gesamtregister von Hausser für jedes Schlagwort nur 0,2 Untereinträge zur Verfügung. Bei Stegmüller sind dies 0,32 Untereinträge pro Schlagwort und bei Cap 0,49. Analog zur ersten Aussage über das Auftreten von Untereinträgen bei den Schlagwörtern verfügen die Schlagwörter

bei Werlens Gesamtregister über die meisten Untereinträge, nämlich pro Schlagwort über 0,57 Untereinträge. Durchschnittlich besitzt jedes 3,04. Schlagwort von allen Registern einen Untereintrag; rechnen wir Linke et al. nicht dazu, so steht im Mittel bei jedem 2,56. Schlagwort ein Untereintrag. Bei sämtlichen Registern verfügt jedes einzelne Schlagwort über 0,33 Untereinträge. Wenn wir erneut das Register von Linke et al. nicht mit einbeziehen, dann steht jedem Schlagwort der anderen vier Gesamtregister 0,39 Untereinträge zur Verfügung. Somit besitzt im Schnitt etwa jedes dritte Schlagwort einen Untereintrag. Wenn wir nur die vier Gesamtregister ohne die Namenregister und ohne das Schlagwortverzeichnis von Linke et al. (das ja keine Untereinträge aufweist) berücksichtigen, steht bei jedem 2,56. Schlagwort ein Untereintrag; 60,94% aller Schlagwörter haben dann keinen präzisierenden Untereintrag.

Im Register von Cap taucht ein einziger Querverweis der Form „siehe“ auf. Hausser benutzt keine Querverweise. In seinem Buchregister sind jedoch Einträge wie „Grammatik, kontextfreie“ und „kontextfreie Grammatik“ zu finden, welche einer Verwendung von Querverweisen nahe kommen. Sowohl Linke et al. wie auch Stegmüller führen je 21 Querverweise auf. Im Register von Werlen werden am meisten Querverweise benutzt, nämlich 25.

Nach Auszählung aller Tokens (also der einzelnen Vorkommen von Wortformen) in den Registern ergeben sich die Zahlenwerte der Tabelle 8. Die Ausdrücke „siehe“ und „siehe auch“ sind nicht erfasst worden.

Anzahl Tokens	Cap	Hausser	Linke at al.	Stegmüller	Werlen	Total
Schlagwörter Namen-register	-	677	-	294	289	1260
Untereinträge Namen-register	-	0	-	0	0	0
Schlagwörter Sachregister	1133	1380	942	1035	1144	5634
Untereinträge Sachregister	445	298	0	615	1157	2515
Schlagwörter Gesamt-register	1133	2057	942	1329	1433	6894
Untereinträge Gesamt-register	445	298	0	615	1157	2515
Total der Tokens	1578	2355	942	1944	2590	9409

Tabelle 8: Anzahl Tokens in den Schlagwörtern und Untereinträgen

Setzt man das Total der Anzahl Tokens in Relation zu der Anzahl Registerseiten, so ist es nicht verwunderlich, dass Werlen die höchste Anzahl Tokens erzielt, ist doch sein Gesamtregister mit 29 Seiten mit Abstand das längste. Dennoch beträgt die Anzahl Tokens pro Registerseite bei Werlen 89,1 Tokens, bei Linke et al., wo das Gesamtregister nur gerade 7 Seiten lang ist, 134,57 Tokens. Dies ist die grösste Anzahl Tokens pro Registerseite der fünf untersuchten Buchregister. Der markante Unterschied ist vor allem auf die unterschiedliche Schriftgrösse zurückzuführen. Ich werde von nun an entsprechend die Anzahl der Schlagwörter und der Untereinträge für Vergleiche der einzelnen Register benutzen und nicht ihre Registerseitenanzahl. In Tabelle 9 ist die Anzahl Tokens mit ihrer Verteilung auf Schlagwörter und Untereinträge der Buchregister aufgeführt.

Anzahl Tokens pro	Cap	Hausser	Linke at al.	Stegmüller	Werlen	Durch-schnitt
Schlagwort Namen-register	-	2,80	-	1,97	1,07	1,95
Untereintrag Namen-register	-	0	-	0	0	0
Schlagwort Sachregister	1,52	1,46	1,36	2,13	1,38	1,57
Untereintrag Sachregister	1,23	1,25	0	3,00	1,84	1,83
Schlagwort Gesamt-register	1,52	1,73	1,36	2,09	1,30	1,60
Untereintrag Gesamt-register	1,23	1,25	0	3,00	1,84	1,83
Eintrag Gesamt-register	2,12	1,98	1,36	3,06	2,36	2,18

Tabelle 9: Tokens pro Schlagwort, Untereintrag und Eintrag

Die Namenregister von Hausser, Stegmüller und Werlen bestehen durchschnittlich aus 1,95 Tokens, was – ausser in ein paar wenigen Fällen – einer Erfassung der Nach- und Vornamen entspricht. Die Sachverzeichnisse weisen bei sämtlichen Büchern eine Anzahl Tokens von mehr als 1 auf, nämlich von 1,57 im Durchschnitt, bei Stegmüller sind es sogar 2,13. Dies bedeutet, dass bei etwas mehr als 50% aller Einträge die Schlagwörter Mehrwortterme sind. Linke et al. verwenden am meisten Einzelwörter, Stegmüller am häufigsten Mehrwortterme. Bei den Untereinträgen variiert die Anzahl Tokens pro Untereintrag stärker als bei den Schlagwörtern. Es werden von 1,23 Tokens pro

Untereintrag bis zu 3,00 Tokens pro Untereintrag verwendet. Etwa 80% aller Untereinträge setzen sich damit aus Mehrworttermen zusammen.

Betrachten wir die Anzahl Tokens pro Eintrag, so ergeben sich bei Linke et al. 1,36 Tokens pro Eintrag, bei Stegmüller die höchste Anzahl von 3,06. Die Werte der anderen drei Register liegen relativ nah beieinander ungefähr in der Mitte der beiden Höchst- und Tiefstwerte. Im Schnitt ergeben sich 2,18 Tokens pro Eintrag, und ein Eintrag konstruiert sich somit in den meisten Fällen aus einem einzelnen Schlagwort und einem einzelnen Wort als dazugehörigem Untereintrag. Das überraschendste Ergebnis, das aus Tabelle 9 hervorgeht, ist die durchschnittliche Anzahl von 1,60 Tokens pro Schlagwort, d.h., dass 60% aller Schlagwörter nicht aus einem einzelnen Wort, sondern aus einem zweigliedrigen Mehrwortterm bestehen. Damit rechtfertigt sich die Verwendung von ganzen Nominal- oder Präpositionalphrasen als Indextermkandidaten, wie wir dies für die Indexierung der Textkörper auch tun werden. Würde die durchschnittliche Anzahl Tokens pro Schlagwort sehr nahe bei 1 liegen, wäre auch eine Verwendung der einzelnen Kopfsubstantive der Phrasen in Frage gekommen. In Abschnitt 5.2.3 werden die einzelnen Bestandteile der Schlagwörter und der Untereinträge im Detail besprochen.

5.2.2 Die Wortkategorien der Einträge

Wir erforschen nun die lexikalischen Wortkategorien der Wortformen, die in unseren fünf Registern auftreten. Die Kategorisierung der Wortformen erfolgte gemäss Gallmann/Sitta: Deutsche Grammatik.[287]

Tabelle 10 listet die Menge der in den fünf Gesamtregistern (d.h. der Schlagwörter sowie Untereinträge) vorkommenden lexikalischen Wortkategorien in absoluten Zahlen auf.

[287] Gallmann/Sitta 1996:24-89

Anzahl	Cap	Hausser	Linke at al.	Stegmüller	Werlen	Total
Substantive	819	952	703	917	1357	4748
Eigennamen	87	711	-	319	303	1420
Adjektive	464	352	188	393	553	1950
Adverbien	7	2	-	3	6	18
Verben (Infinitive)	-	-	3	9	2	14
Ziffern	17	3	-	1	-	21
römische Zahlen	-	2	-	-	-	2
bestimmte Artikel	13	13	4	107	153	290
unbestimmte Artikel	14	-	-	17	1	32
Indefinit-pronomen	2	-	-	4	-	6
Possessiv-pronomen	-	-	-	1	-	1
Präpositionen	62	9	5	81	82	239
beiordnende Konjunktionen	1	6	5	32	28	72
unterordnende Konjunktionen	1	-	-	-	-	1
Partikel „als“	3	-	-	24	10	37
englische Wortformen	46	183	25	-	59	313
lateinische Wortformen	6	4	-	29	11	50
französische Wortformen	-	1	7	-	22	30
andere fremd-sprachige Wortformen	-	-	-	-	2	2
Akronyme/ Symbole/ Sonderzeichen	36	117	2	7	1	163
Total Wortformen	1578	2355	942	1944	2590	9409

Tabelle 10: Lexikalische Wortkategorien in den fünf Registern

Die gesamte Tabelle 10 in Prozenten ausgedrückt ergibt die Werte der nachfolgenden Tabelle 11. Damit lassen sich vergleichende Aussagen besser machen. Sämtliche in den fünf Registern auftretenden Wortkategorien sind darin aufgeführt. Die Wortklassenproportionen der einzelnen Autoren wurden bezüglich der Gesamtanzahl Tokens der einzelnen Buchregister berechnet.

Anteil in %	Cap	Hausser	Linke at al.	Stegmüller	Werlen	Total
Substantive	51,90%	40,42%	74,63%	47,17%	52,39%	50,46%
Eigennamen	5,51%	30,19%	-	16,41%	1,70%	15,09%
Adjektive	29,40%	14,95%	19,96%	20,22%	21,35%	20,72%
Adverbien	0,44%	0,08%	-	0,15%	0,23%	0,19%
Verben (Infinitive)	-	-	0,32%	0,46%	0,08%	0,15%
Ziffern	1,08%	0,13%	-	0,05%	-	0,22%
römische Zahlen	-	0,08%	-	-	-	0,02%
bestimmte Artikel	0,82%	0,55%	0,42%	5,50%	5,91%	3,08%
unbestimmte Artikel	0,89%	-	-	0,87%	0,04%	0,34%
Indefinit-pronomen	0,13%	-	-	0,21%	-	0,06%
Possessiv-pronomen	-	-	-	0,05%	-	0,01%
Präpositionen	3,93%	0,38%	0,53%	4,17%	3,17%	2,54%
beiordnende Konjunktionen	0,06%	0,25%	053%	1,65%	1,08%	0,77%
unterordnende Konjunktionen	0,06%	-	-	-	-	0,01%
Partikel „als“	0,19%	-	-	1,23%	0,39%	0,39%
englische Wortformen	2,92%	7,77%	2,65%	-	2,28%	3,33%
lateinische Wortformen	0,38%	0,17%	-	1,49%	0,42%	0,53%
französische Wortformen	-	0,04%	0,74%	-	0,85%	0,32%
andere fremd-sprachige Wortformen	-	-	-	-	0,08%	0,02%
Akronyme/ Symbole/ Sonderzeichen	2,28%	4,97%	0,21%	0,36%	0,04%	1,73%
Total Wortformen	100%	100%	100%	100%	100%	100%

Tabelle 11: Lexikalische Wortkategorien in Prozent zum Total der Tokens

50,46% aller Wortformen in den Indexen sind Substantive, bei Linke et al. erreicht die Klasse der Substantive sogar einen Anteil von 74,63%. Auch im schlechtesten Fall von Hausser beträgt der Anteil der Substantive immerhin noch 40,42%. Zusätzlich erzielen die Eigennamen im Schnitt einen Wert von 15,09%. Damit bestätigt sich die Aussage von Mulvany, dass sich insbesondere die Substantive als Indexterme eignen (vergleiche Abschnitt 3.2.1.1). Ob die Substantive im Speziellen für die Ebene der Schlagwörter geeignet sind, erfahren wir in Abschnitt 5.2.3. Die Klasse der Adjektive ist mit einem Anteil von 20,72% aller Wortformen vertreten. Der Anteilsbereich erstreckt sich von 14,95% bei Hausser bis zu 29,40% bei Cap.

Die Portionen aller anderen Wortarten betragen weniger als 3,5%. Für aussagekräftigere Interpretationen der Zahlenwerte werden die einzelnen Kategorien deshalb zusammengefasst, und zwar wie folgt: Substantive und Eigennamen ergeben zusammen die Klasse *Substantive*. Sämtliche Adjektive, Adverbien und Zahlausdrücke werden zu einer einzigen Klasse *Adjektive* zusammengefasst. Die Verben konstituieren eine eigene Kategorie. Die definiten und indefiniten Artikel, alle Pronomen, die Präpositionen, alle Konjunktionen sowie die übrigen Partikeln formieren die Klasse der Funktionswörter. Die fremdsprachigen Ausdrücke werden nicht mehr unterteilt, sondern als eine einzige Kategorie *fremdsprachige Ausdrücke* erfasst. Die Akronyme, Abkürzungen, Symbole und Sonderzeichen bleiben als Klasse *spezielle Ausdrücke* erhalten. Aus diesen Zusammenziehungen resultiert die Tabelle 12, die in absoluten Zahlen und jeweils darunter stehend in Prozentzahlen die Anteile der neuen zusammengefassten Wortkategorien ausdrückt.

Anteil	Cap	Hausser	Linke at al.	Stegmüller	Werlen	Total
Substantive	906	1663	703	1236	1660	6168
	57,41%	70,62%	74,63%	63,58%	64,09%	65,55%
Adjektive	488	359	188	397	559	1991
	30,93%	15,24%	19,96%	20,42%	21,58%	21,16%
Verben	-	-	3	9	2	14
	-	-	0,32%	0,46%	0,08%	0,15%
Funktionswörter	96	28	14	266	274	678
	6,08%	1,19%	1,49%	13,68%	10,58%	7,21%
fremdsprachige Ausdrücke	52	188	32	29	94	395
	3,30%	7,98%	3,40%	1,49%	3,63%	4,20%
spezielle Ausdrücke	36	117	2	7	1	163
	2,28%	4,97%	0,21%	0,36%	0,04%	1,73%
Total	1578	2355	942	1944	2590	9409
	100%	100%	100%	100%	100%	100%

Tabelle 12: Wortkategorien zusammengefasst 1

Aus dieser Tabelle ist nun auf einfachere Weise ersichtlich, dass der Anteil aller Substantive 65,55% der Wortformen in den fünf untersuchten Registern beträgt. Der Adjektivklassenanteil hat sich nur leicht auf 21,16% erhöht. Die Menge der Funktionswörter macht 7,21% aller Wortformen aus, und die fremdsprachigen Wörter erzielen eine erstaunlich hohe Portion von 4,20%. Mit 1,73% Anteil rangieren die speziellen Ausdrücke an zweitletzter sowie die Verben mit 0,15% an letzter Stelle. Dies alles bedeutet, dass bei einer automatischen Indexierungsmethode mit einer Extraktion der Nominal- und Präpositionalphrasen aus den Textkörpern und mit einer Verwendung einer Stoppwortliste ohne Gewichtung der Terme 13,29% aller in den Buchindexen vorkommenden relevanten Wortformen verfehlt, 86,71% jedoch erwischt würden, sofern die Autoren sowohl für die Buchtexte wie für die Register dieselben Ausdrucksweisen verwenden würden. Der Anteil an Ballast wäre allerdings unausgesprochen hoch.

Fassen wir die Klassen der Tabelle 12 noch weiter zusammen, so ergeben sich die Wortkategorieklassen *sinntragende Wörter* (Substantive, Adjektive und Verben), *Funktionswörter* und *spezielle Ausdrücke* (fremdsprachige und spezielle Ausdrücke). In Tabelle 13 sind die Verhältnisse der drei Kategorien dargestellt.

Anteil	Cap	Hausser	Linke at al.	Stegmüller	Werlen	Total
sinntragende Wörter	1394 88,34%	2022 85,86%	894 94,90%	1642 84,47%	2221 85,75%	8173 86,86%
Funktionswörter	96 6,08%	28 1,19%	14 1,49%	266 13,68%	274 10,58%	678 7,21%
spezielle Ausdrücke	88 5,58%	305 12,95%	34 3,61%	36 1,85%	95 3,67%	558 5,93%
Total	1578 100%	2355 100%	942 100%	1944 100%	2590 100%	9409 100%

Tabelle 13: Wortkategorien zusammengefasst 2

Die sinntragenden Wörter in den fünf Buchregistern machen 86,86% aller Wörter aus. Dabei ist zu beachten, dass der Anteil von 0,15% an Verben bei einer Indexierung mithilfe von Nominal- und Präpositionalphrasenextraktion nicht als Indextermkandidatenbestandteile zum Zuge kommen würden. Ob die speziellen Ausdrücke mögliche Indexterme darstellen, würde in erster Linie vom Extraktionswerkzeug abhängen. Wie wir indessen aus der Übersicht ersehen können, fällt der Anteil der speziellen Ausdrücke bei Hausser (12,95%) und vielleicht noch bei Cap (5,58%) ins Gewicht.

Die Eliminierung der Funktionswörter durch Anwendung einer Stoppwortliste würde einen Anteil von 7,21% der Wortformen in den Indexen entfernen. Bei der Indexierung der Textkörper in Abschnitt 5.3 werden trotz all dieser Aussagen über das Verpassen von Wortformen nur die Nominal- und Präpositionalphrasen aus den Texten gefiltert und die Funktionswörter eliminiert, obschon bei Stegmüller und bei Werlen der Anteil der Funktionswörter mit 13,68% und 10,58% recht beträchtlich ist. Durch manuelle Nachprüfung der extrahierten Phrasen wurde sichergestellt, dass die fremdsprachigen und speziellen Ausdrücke als mögliche Indextermkandidaten Verwendung fanden, sofern diese innerhalb einer Nominal- oder Präpositionalphrase auftraten.

Mithilfe der folgenden Tabelle 14 möchte ich nun erforschen, wie sich die lexikalischen Wortkategorien auf die Schlagwörter und Untereinträge verteilen.

Anteil	Cap	Hausser	Linke at al.	Stegmüller	Werlen	Total
Substantive	784	1517	703	918	1137	5059
Schlagwörter	49,68%	64,42%	74,63%	47,22%	43,90%	53,77%
Substantive	122	146	-	318	523	1109
Untereinträge	7,73%	6,20%	-	16,36%	20,19%	11,79%
Adjektive	240	272	188	247	156	1103
Schlagwörter	15,21%	11,55%	19,96%	12,71%	6,02%	11,72%
Adjektive	248	87	-	150	403	888
Untereinträge	15,72%	3,69%	-	7,72%	15,56%	9,44%
Verben	-	-	3	1	-	4
Schlagwörter	-	-	0,32%	0,05%	-	0,04%
Verben	-	-	-	8	2	10
Untereinträge	-	-	-	0,41%	0,08%	0,11%
Funktionswörter	52	24	14	146	50	286
Schlagwörter	3,30%	1,02%	1,49%	7,51%	1,93%	3,04%
Funktionswörter	44	4	-	120	224	392
Untereinträge	2,79%	0,17%	-	6,17%	8,65%	4,17%
fremdsprachige Wörter	33	182	32	13	90	350
Schlagwörter	2,09%	7,73%	3,40%	0,67%	3,47%	3,72%
fremdsprachige Wörter	19	6	-	16	4	45
Untereinträge	1,20%	0,25%	-	0,82%	0,15%	0,49%
spezielle Ausdrücke	24	62	2	4	-	92
Schlagwörter	1,52%	2,63%	0,21%	0,21%	-	0,98%
spezielle Ausdrücke	12	55	-	3	1	71
Untereinträge	0,76%	2,34%	-	0,15%	0,04%	0,75%
Total	1133	2057	942	1329	1433	6894
Schlagwörter	71,80%	87,35%	100%	68,36%	55,33%	73,27%
Total	445	298	-	615	1157	2515
Untereinträge	28,20%	12,65%	-	31,64%	44,67%	26,73%
Gesamttotal	1578	2355	942	1944	2590	9409
	100%	100%	100%	100%	100%	100%

Tabelle 14: Wortkategorieklassen in den Schlagwörtern und Untereinträgen

Aus der Aufstellung geht hervor, dass 53,77% aller Wortformen als Substantive (inklusive Eigennamen) in den Schlagwörtern erscheinen, gegenüber 11,79% Substantiven in den Untereinträgen. Das wiederum heisst, dass 82,0% aller in den fünf Indexen vorkommenden Substantive in den Schlagwörtern auftreten. Allerdings ist hier einschränkend zu bemerken, dass der gesamte Anteil der Substantive bei Linke et al. nur auf der Ebene der Schlagwörter auftritt, da dieses Register keine Untereinträge besitzt.

Bei den Adjektiven ist die Verteilung auf Schlagwörter und Untereinträge recht ausgeglichen: 11,72% der Wortformen sind in den Schlagwörtern stehende Adjektive, 9,44% beträgt der Anteil der Adjektivformen, die in den Untereinträgen erscheinen. Ebenso sind die Funktionswörter, die speziellen Ausdrücke sowie die Verben fast zu gleichen Teilen auf die Schlagwörter und die Untereinträge verteilt. Bei den fremdsprachigen Wortformen zeigt sich, dass 88,61% aller Fremdsprachenausdrücke auf der Ebene der Schlagwörter auftreten und nur 11,39% auf der Untereintragebene. Bei der Verarbeitung von deutschen Fachtexten scheint es somit angebracht, auch die fremdsprachigen Ausdrücke mit einzubeziehen, da ein beträchtlicher Anteil derer als Schlagwort oder Schlagwortbestandteil zu finden ist. Insbesondere beim Register von Hausser machen die fremdsprachigen Wortformen, die in den Schlagwörtern auftreten, einen nicht zu übersehbaren Anteil von 7,73% aus. In Anlehnung an diese Beobachtung wird die Selektion der fremdsprachigen Nominal- und Präpositionalphrasen der untersuchten Textkörper als Indextermkandidaten erneut bestätigt.

Auf eine Unterscheidung von Substantiven, die in einem Schlagwort als Ordnungswort fungieren, und Substantiven, welche Bestandteil eines Schlagwortes sind, jedoch nicht dessen Ordnungswort, wurde verzichtet. Denn bei der Extraktion der Nominal- und Präpositionalphrasen aus den fünf Textkörpern wurden die Kopfsubstantive einer Phrase immer als Ordnungswort an die erste Stelle eines Indextermkandidaten gestellt.

Die nachstehende Tabelle 15 fasst die Wortkategorien in Bezug auf ihre Verteilung auf Schlagwörter und Untereinträge nochmals auf die drei Hauptkategorien (sinntragende Wörter, Funktionswörter, spezielle Ausdrücke) zusammen.

Anteil	Cap	Hausser	Linke at al.	Stegmüller	Werlen	Total
sinntragende Wörter	1024	1789	894	1166	1293	6166
Schlagwörter	64,89%	75,97%	94,90%	59,98%	49,92%	65,53%
sinntragende Wörter	370	233	-	476	928	2007
Untereinträge	23,45%	9,89%	-	24,49%	35,83%	21,33%
Funktionswörter	52	24	14	146	50	286
Schlagwörter	3,30%	1,02%	1,49%	7,51%	1,93%	3,04%
Funktionswörter	44	4	-	120	224	392
Untereinträge	2,79%	0,17%	-	6,17%	8,65%	4,17%
spezielle Ausdrücke	57	244	34	17	90	442
Schlagwörter	3,61%	10,36%	3,61%	0,87%	3,47%	4,70%
spezielle Ausdrücke	31	61	-	19	5	116
Untereinträge	1,96%	2,59%	-	0,98%	0,19%	1,23%
Total	1133	2057	942	1329	1433	6894
Schlagwörter	71,80%	87,35%	100%	68,36%	55,33%	73,27%
Total	445	298	-	615	1157	2515
Untereinträge	28,20%	12,65%	-	31,64%	44,67%	26,73%
Gesamttotal	1578	2355	942	1944	2590	9409
	100%	100%	100%	100%	100%	100%

Tabelle 15: Wortkategorieklassen zusammengefasst in den Schlagwörtern und Untereinträgen

Die sinntragenden Wörter verteilen sich nun mit 65,53% auf die Schlagwörter und mit 21,33% auf die Untereinträge. Oder anders formuliert stehen 75,44% aller sinntragenden Wörter bei unseren fünf Registern in den Schlagwörtern. Die Funktionswörter treten etwas häufiger in den Untereinträgen (4,17%) als in den Schlagwörtern (3,04%) auf. Dies ist nicht erstaunlich, denn Untereinträge können sehr frei und vielfältig formuliert werden und Phrasen beschreiben Sachverhalte präziser als einzelne Wörter. Bei den speziellen Ausdrücken erreichen diejenigen, die auf der Schlagwortebene erscheinen, mit 4,70% einen gering höheren Anteil als diejenigen auf der Untereintragebene mit 1,23%.

Dass die Substantive einen grossen und wichtigen Teil der Indexwortformen ausmachen, ist evident geworden. Wir wollen deshalb die Substantive genauer betrachten und uns einen Überblick über die

Proportionen innerhalb der Klasse der Substantive verschaffen. Tabelle 16 gibt Auskunft über die Anteile der Kompositasubstantive, der Pluralsubstantivformen und der Singularsubstantivformen.

Anteil	Cap	Hausser	Linke at al.	Stegmüller	Werlen	Total
Komposita	275	376	268	216	326	1461
	30,35%	22,61%	38,12%	17,48%	19,64%	23,69%
Pluralformen	65	114	9	136	148	472
	7,17%	6,86%	1,28%	11,00%	8,92%	7,65%
Singularformen	841	1549	694	1100	1512	5696
	92,83%	93,14%	98,72%	89,00%	91,08%	92,35%
Gesamttotal	906	1663	703	1236	1660	6168
der Substantive	100%	100%	100%	100%	100%	100%

Tabelle 16: Substantive: Komposita, Plural- und Singularformen

Es ist nicht verwunderlich, dass der kürzeste Index von Linke et al. signifikant mehr Komposita gebraucht als die anderen vier Indexe, dienen doch gerade die Komposita der Formulierungskomprimierung. 38,12% aller Substantive bei Linke et al. sind Komposita, bei Stegmüller nur gerade 17,48%. Im Schnitt sind bei allen fünf Buchregistern 23,69% aller Substantive Komposita im Sinne von z.B. „Apperzeptionsorgan“. Mit einer Begriffsanalyse bzw. Kompositazerlegung wäre es möglich, auch über die einzelnen Bestandteile eines Kompositums als Indextermkandidaten zu verfügen. Ich werde auf eine Kompositaanalyse innerhalb der Prüfung der inversen Seitenhäufigkeit dennoch verzichten, da die in den Textkörpern vorkommenden Komposita meiner Meinung nach wesentlich mehr Ballast als Nutzen bringen werden. Ausserdem werden diejenigen Komposita in den Registern, welche Fachtermini bilden, mit grosser Wahrscheinlichkeit auch als Komposita in den Texten auftreten.

Bei allen fünf Büchern kommen in den Registern sowohl Singular- wie auch Pluralformen von Substantiven vor. Der Anteil der Pluralsubstantivformen ist jedoch mit durchschnittlich 7,65% eher gering; bei Linke et al. erreichen die Pluralformen sogar nur eine Portion von 1,28%, im Index von Stegmüller sind es immerhin 11,0%. Die Einhaltung einer sturen Konsistenz bei der Verwendung von Plural- oder Singularformen wird in keinem der Register verfolgt. Vielmehr wird bei einer Betrachtung der einzelnen Pluralformen klar, dass die Pluralsubstantive gezielt verwendet werden und zuweilen eine Verwendung beider Formen gewählt wird, um deren unterschiedliche Bedeutungen zu erfassen. So erscheint bei Linke et al. beispielsweise das Schlagwort „Kardinalvokale“, während fast alle anderen Schlagwortsubstantive im Singular stehen. Werlen führt sowohl „Farbwort“ wie auch „Farbwörter“ als Schlagwörter auf. Durch die Anwendung der Lemmatisierungstechnik bei den extrahierten Wortgruppen zur Überprüfung der inversen

Seitenfrequenz geht somit eine gewisse Anzahl Substantivschlagwörter oder Substantivuntereinträge verloren. Dennoch werden die Wortformen der Textkörper lemmatisiert und auf die Singularformen reduziert, um möglichst viele Ausdrucksweisen eines Indextermkandidaten auf eine einheitliche Form zu bringen und ein möglichst hohes dazugehöriges Gewicht der inversen Seitenfrequenz zu erhalten. Der Grund dafür ist, dass ich das Aufführen eines Schlagwortes im Singular einem gänzlichen Nichtvorhandensein des Schlagwortes in einem Buchregister vorziehe.

5.2.3 Die Bestandteile der Einträge

In diesem Unterkapitel zur Untersuchung der Schlagwortregister wird ermittelt, aus welchen Bestandteilen bzw. aus welchen lexikalischen Wortkategorien oder syntaktischen Konstituenten sich die Einträge in den fünf untersuchten Registern zusammensetzen. Wir werden dabei nicht mehr auf die Mengenverhältnisse der einzelnen Wortklassen gesamthaft, sondern auf die Quantitäten der Wortkategorien in Bezug auf die einzelnen Einträge eingehen. Das Ziel ist, herauszufinden, wie ein „typischer“ Eintrag aussieht, aus welchen lexikalischen und syntaktischen Bestandteilen er gebildet ist.

Zunächst wollen wir uns einen Überblick darüber verschaffen, aus welchen Bestandteilen die Schlagwörter (ohne Untereinträge) in unseren fünf Buchindexen bestehen. Die auf den nächsten fünf Seiten aufgeführte Tabelle 17 stellt diese Elemente in absoluten Zahlen sowie Prozentzahlen dar. Die Ausdrücke „siehe“ und „siehe auch“ sind nicht berücksichtigt worden. Sämtliche Adverbien wurden als Adjektive gezählt. Eckige Klammern drücken Zusammengehörigkeit aus, runde Klammern Optionalität.

Zusammensetzung Schlagwort	Beispiel(e)	Cap	Hausser	Linke et al.	Stegmüller	Werlen	Total
Substantiv	Addierer	385 51,75%	586 49,24%	512 73,88%	187 29,40%	847 77,07%	2517 57,70%
Substantiv + Substantiv	Ackermann Funktion	52 6,99%	33 2,77%	3 0,43%	5 0,79%	21 1,91%	114 2,61%
Substantiv + Substantiv + Substantiv	Average Case Analysis	14 1,88%	15 1,26%	1 0,14%		1 0,09%	31 0,71%
Substantiv + Substantiv als nähere Bestimmung	Komplexität, Sortierverfahren	11 1,48%	246 20,67%	3 0,43%	150 23,58%	19 1,73%	429 9,83%
Substantiv + [Adjektiv + Substantiv als nähere Bestimmung]	Intension, höhere Ordnung		8 0,67%		1 0,16%	4 0,36%	13 0,30%
Substantiv + „und“ + Substantiv	Prinzip und Parameter	3 0,40%		1 0,14%	2 0,31%		6 0,14%
Substantiv + „als“ + (Adjektiv) + Substantiv	Teilmengenrelation als Ordnungs-relation Erklärung als mehr-fache Analyse	1 0,13%			8 1,26%		9 0,21%
Substantiv + „als“ + Substantiv + (Genitivattribut oder Präpositionalphrase)	Extension als Auswertung einer Intension Falsifizierbarkeit als Kriterium für Nichtnormativität				2 0,31%		2 0,05%

Zusammensetzung Schlagwort	Beispiel(e)	Cap	Hausser	Linke et al.	Stegmüller	Werlen	Total
Substantiv + ein Adjektiv nachgestellt	Äquivalenz, asymptotische	12 1,61%	3 0,25%	135 19,48%	39 6,13%	86 7,83%	275 6,30%
Substantiv + zwei Adjektive nachgestellt, durch Konjunktion oder Präposition verknüpft	Merkmal, phonologisch distinktives Fertigkeit, produktive und rezeptive Phonologie, lineare vs. nichtlineare			6 0,87%	4 0,63%	1 0,09%	11 0,25%
Substantiv + ein Adjektiv nachgestellt + nähere Bestimmung	Verarbeitung, sprachliche des vorsprachlichen Denkens Eigenart, geistige von Völkern Widerspruch, logischer und Realkonflikt				1 0,16%	4 0,36%	5 0,11%
ein Adjektiv vorangestellt + Substantiv	algebraische Semantik	139 18,68%	248 20,84%		101 15,88%	18 1,64%	506 11,60%
zwei Adjektive vorangestellt + Substantiv	linear beschränkter Automat	6 0,81%	9 0,76%		5 0,79%		20 0,46%

Zusammensetzung Schlagwort	Beispiel(e)	Cap	Hausser	Linke et al.	Stegmüller	Werlen	Total
ein Adjektiv voran-gestellt + Substantiv + irgendeine nähere Bestimmung	polnische Notation, umgekehrte empirische Äquivalenz unverträglicher Theorien hypothetische Rechtfertigung von Handlungen lingualistische Lehre von der logischen Wahrheit	4 0,54%			13 2,04%		17 0,39%
Substantiv + einfaches Genitivattribut	Anwendung einer Regel	9 1,21%	10 0,84%	1 0,14%	12 1,89%	15 1,36%	47 1,08%
Substantiv + komplexes Genitivattribut	Desiderata der generativen Grammatik		3 0,25%		9 1,42%	3 0,27%	15 0,34%
Substantiv + (Substantiv + Artikel als nähere Bestimmung)	Exportation, Tautologie der Bedeutung, abstrakte Interpretation der	3 0,40%		2 0,29%	16 2,52%	12 1,09%	33 0,76%

Zusammensetzung Schlagwort	Beispiel(e)	Cap	Hausser	Linke et al.	Stegmüller	Werlen	Total
Substantiv + komplexes Genitiv-attribut (als nähere Bestimmung)	Sitte, Ausdruck der Volksseele Wurzel, erste Bestandteile der Sprache Lösung des Problems der theoretischen Begriffe				7 1,10%	5 0,45%	12 0,28%
Substantiv + einfache Präpositionalphrase	Anzahl von Elementen	12 1,61%	4 0,34%	2 0,29%	15 2,36%	6 0,55%	39 0,89%
Substantiv + komplexe Präpositionalphrase	Theorem von Rice und Shapiro Präferenzordnung über der Gesamtheit aller möglichen Welten Hinderungsgründe für die Erfüllung von Verpflichtungen	1 0,13%	2 0,17%		16 2,5%		19 0,44%
Substantiv + irgendeine Art von Präpositionalphrase als nähere Bestimmung	Asser, Tautologie von Phonetik, Abgrenzung zur Phonologie Kausalprozesse, zwei Arten von	17 2,28%	2 0,17%	4 0,58%	12 1,89%	9 0,82%	44 1,01%

Zusammensetzung Schlagwort	Beispiel(e)	Cap	Hausser	Linke et al.	Stegmüller	Werlen	Total
Adjektiv	adjazent möglicherweise	44 5,91%	14 1,18%	16 2,31%	12 1,89%	43 3,91%	129 2,96%
Adjektiv + Adjektiv	explizit performativ	9 1,21%	5 0,42%	2 0,29%	4 0,63%	1 0,09%	21 0,48%
Adjektiv + Adjektiv + Adjektiv	monoton rekursiv aufzählbar	1 0,13%					1 0,02%
Adjektiv + Präpositionalphrase	intensional bezüglich einer Beschreibung				4 0,63%		4 0,09%
Verb (Infinitiv)	klassifizieren	2 0,27%		3 0,43%			5 0,11%
Konjunktion	and	2 0,27%					2 0,05%
Akronyme/Abkürzungen/Symbole	DES L_2	12 1,61%			4 0,63%		16 0,37%
Andere	genau ein, Gegensatz zu ein	2 0,27%			7 1,10%	4 0,36%	14 0,32%
nicht gezählte	ex falso quodlibet	3 0,40%	2 0,17%	2 0,29%			6 0,14%
Total		744 100%	1190 100%	693 100%	636 100%	1099 100%	4362 100%

Tabelle 17: Zusammensetzung der Schlagwörter

Ganze Teilsätze (z.B. „how to say it“, „ex falso quodlibet“) wurden in der Tabelle 17 als „nicht gezählte“ erfasst. Die Schlagwörter, welche unter „andere“ aufgeführt wurden, sind die folgenden Schlagwörter (sie passen in keines der anderen Schemas und kommen jeweils nur einmal vor):

drei, Zahl
ein, Gegensatz zu genau ein
Erklären, als Freilegen von Details
Fregesches Fundamentalprinzip in der Montague-Grammatik, erstes
genau ein, Gegensatz zu ein
Identitätssätze als notwendige Wahrheiten a posteriori
Parallelisierung, Sprache und Kultur
Präsuppositionen, unerfüllte als Motiv für eine dreiwertige Logik
Tarskischer Wahrheitsbegriff als metaphysisch und erkenntnistheoretisch neutral
Tier, Mensch: Unterschied
unter Konkatenation abgeschlossen
Verba im Präteritum, als ursprüngliche Wörter
weil als Aussagenverbindung

Die grösste Anzahl der Schlagwörter, 57,7%, besteht aus einem einzigen Substantiv. Beim Index von Werlen stellen die allein stehenden Substantive sogar einen Anteil von 77,07%, der geringste Anteil bei Stegmüller macht 29,40% aus. Die Schlagwörter, die aus zwei oder drei Substantiven zusammengesetzt sind, erreichen 3,32% aller Schlagwörter. Am zweithäufigsten sind in allen Registern die Schlagwörter, die aus einem vorangestellten Adjektiv als Ordnungswort und einem Substantiv bestehen. Bei Hausser erreicht diese Kategorie 20,84%, bei Werlen allerdings nur 1,64%. Im Durchschnitt machen sie einen 11,6-prozentigen Anteil aus. Bei den aus einem Substantiv und einem Substantiv als nähere Bestimmung konstruierten Schlagwörtern beträgt der Anteil in allen Registern 9,83%. Der Index von Hausser erzielt 20,67%, Stegmüller sogar 23,58%. Diese hohen Portionen sind auf die Namenindexe zurückzuführen, die aus den Nachnamen als Ordnungswörtern sowie aus den Vornamen als nähere Bestimmungen bestehen. Bei Werlen kommen im Namenindex in den allermeisten Fällen nur die Nachnamen vor, sodass sich dieser Teil bei den allein stehenden Substantiven niederschlägt und den hohen Prozentsatz von 77,07% ausmacht. Die Bücher von Cap und Linke et al., welche keine Namenregister besitzen, weisen in ihren Sachregistern nur sehr kleine Anteile von Schlagwörtern auf, die aus einem Substantiv und einem Substantiv als nähere Bestimmung bestehen, nämlich 1,48% bei Cap und 0,43% bei Linke et al.

Nur gerade 2,96% aller Schlagwörter bestehen aus einem einzelnen Adjektiv. Der Index von Cap erzielt dabei den höchsten Anteil mit 5,91%. Zählen wir diese Schlagwörter sowie diejenigen Schlagwörter zusammen, die sich aus einem Adjektiv als Ordnungswort und irgendeiner zusätzlichen Angabe zusammensetzen, jedoch nicht aus einem einzelnen Adjektiv, so erhalten wir einen Anteil von 3,55% aller Schlagwörter in allen Registern. Die Adjektive spielen als Ordnungswörter auf der Ebene der Schlagwörter also eine eher unbedeutende Rolle. Das Gleiche

gilt für die infiniten Verben (0,11%), Konjunktionen (0,05%) und die Kategorie der Akronyme/Abkürzungen/Symbole (0,37%).

Schlagwörter, bestehend aus einem Substantiv und einem oder zwei Adjektiven, die voran- oder nachgestellt sind, erzielen bei den fünf Registern einen Anteil von 18,61%, während die Substantive, die durch ein Genitivattribut oder eine Präpositionalphrase präzisiert werden, nur einen geringen Teil der Schlagwörter ausmachen (4,04%). Wenn wir alle Schlagwörter, die als Ordnungswort ein Substantiv besitzen und irgendeine nachgestellte Form von näherer Bestimmung, z.B. Adjektiv, „als“-Ausdruck, Präpositionalphrase usw., so beträgt deren Anteil 20,99% aller Schlagwörter. Nehmen wir auch die allein stehenden Substantive und diejenigen, die aus zwei oder drei Substantiven gebildet sind, hinzu, so misst der Anteil der Schlagwörter, die ein Substantiv als Ordnungswort aufweisen, 83,01%. Kehren wir nun noch die Reihenfolge der aus einem oder zwei vorangestellten Adjektiven und einem Substantiv bestehenden Schlagwörter um, sodass sämtliche Adjektive nachgestellt sind, so machen die Schlagwörter, die sich aus einem Substantiv als Ordnungswort und irgendwelchen zusätzlichen Angaben zusammensetzen, den äusserst grossen Anteil von 95,46% der Schlagwörter in allen Registern aus.

Aus Tabelle 17 geht auch hervor, dass in allen fünf Gesamtregistern 38,81% aller Schlagwörter Multiwortterme sind und 61,19% einzelne Wörter. Die längsten gefundenen Schlagwörter bestehen aus acht Tokens und lauten: „Präsuppositionen, unerfüllte als Motiv für eine dreiwertige Logik“, „Unbestimmtheit aller Theorien über die Natur, These der“. Die häufigste Kombination der Multiwortterme ist: ein Substantiv und ein vor- oder nachgestelltes Adjektiv. Diese Feststellung bestätigt, dass es durchaus sinnvoll ist, beim automatischen Extraktionsindexieren nicht nur einzelne Wörter, sondern ganze Phrasen zu selektieren und als Indextermkandidaten zu verwenden, insbesondere Nominalphrasen mit eingebetteten Adjektivphrasen.

Nach der Analyse der Schlagwörter werden wir nun auch die lexikalischen und syntaktischen Elemente der Untereinträge betrachten. Diese sind analog zur Tabelle der Schlagwortbestandteile in Tabelle 18 aufgelistet. Linke et al. werden nicht aufgeführt, da deren Register keine Untereinträge besitzt.

Zusammensetzung Untereintrag	Beispiel(e)	Cap	Hausser	Stegmüller	Werlen	Total
Substantiv	Programme	102 28,25%	113 47,48%	42 20,49%	114 18,10%	371 25,87%
Substantiv + Substantiv	Meally Automat	3 0,83%			1 0,16%	4 0,28%
Substantiv + Substantiv als nähere Bestimmung	Konsonanten, Artikulation		4 1,68%		5 0,79%	9 0,63%
Substantiv + Artikel	Einheit des Theorien der	2 0,55%		21 10,24%	39 6,19%	62 4,32%
Substantiv + Präposition	Wörter für	6 1,66%		2 0,98%	5 0,79%	13 0,91%
Substantiv + (Adjektiv + Substantiv als nähere Bestimmung)	Denken, gegenseitiger Einfluss				4 0,63%	4 0,28%
Substantiv + (Substantiv + Artikel als nähere Bestimmung)	Funktionalität, Grundsatz der			1 0,49%		1 0,07%
Substantiv + „und“ + Substantiv	Prinzip und Parameter			10 4,88%	6 0,95%	16 1,12%
„und“ + Substantiv	und Energeia				6 0,95%	6 0,42%
„als“ + Substantiv	als Relation	1 0,28%		1 0,49%		2 0,14%
„als“ (+ Adjektiv) + Substantiv + (n.B.)	als Wissenschaft der Noemata			4 1,95%	8 1,27%	12 0,84%

Zusammensetzung Untereintrag	Beispiel(e)	Cap	Hausser	Stegmüller	Werlen	Total
Substantiv + ein Adjektiv nachgestellt	Typ 1	18 4,99%	4 1,68%		1 0,16%	23 1,60%
ein Adjektiv vorangestellt + Substantiv	deklarative Aspekte	4 1,11%	13 5,46%	7 3,41%	80 12,70%	104 7,25%
ein Adjektiv vorangestellt + Substantiv + Artikel	psychologische Interpretation der			9 4,39%		9 0,63%
ein Adjektiv vorangestellt + Substantiv + irgendeine nähere Bestimmung	bildendes Organ des Gedankens kaleidoskopischer Strom von Eindrücken				6 0,95%	6 0,42%
zwei Adjektive voran-gestellt (evtl. mit „und“ verknüpft) + Substantiv	wissenschaftliche, symbolische Form eingeborene und sprachliche Relativität				6 0,95%	6 0,42%
„und“ + Adjektiv voran-gestellt + Substantiv	und opake Kontexte			5 2,44%		5 0,35%
einfaches Genitivattribut	einer Tabelle	5 1,36%		5 2,44%	10 1,59%	20 1,39%
komplexes Genitivattribut	der positiven ganzen Zahlen höherer Ordnung des Glaubens und Wissens Quines Charakterisierung	9 2,49%	1 0,42%	9 4,39%	6 0,95%	25 1,74%
Substantiv + einfaches Genitivattribut	Philosophie Putnams Masse des Denkbaren		1 0,42%	1 0,49%	29 4,60%	31 2,16%

Zusammensetzung Untereintrag	Beispiel(e)	Cap	Hausser	Stegmüller	Werlen	Total
Substantiv + komplexe Art von Genitivattribut als nähere Bestimmung	Bedingung der Möglichkeit des Philosophierens Austausch physischer Zeichen				7 1,11%	7 0,49%
einfache Präpositionalphrase	von Relationen in der Montague-Grammatik	14 3,88%	1 0,42%	8 3,90%	10 1,59%	33 2,30%
komplexe Präpositionalphrase	durch primitive Rekursion mit Vorzug der Freiheit im Sinne von Kuhn vor und nach Humboldt	1 0,28%	2 0,84%	6 2,93%	10 1,59%	19 1,32%
Substantiv + einfache Präpositionalphrase	Definition durch Grammatik	1 0,28%	1 0,42%		28 4,44%	30 2,09%
Adjektiv	syntaktische	174 48,20%	68 28,57%	56 27,32%	207 32,86%	505 35,22%
Adjektiv + Adjektiv [288]	simultane induktive formlos, gestaltlos	12 3,32%		2 0,98%	16 2,54%	30 2,09%
Adjektiv + Adjektiv + Adjektiv	monoton rekursiv aufzählbare	1 0,28%				1 0,07%
Adjektiv + „und“ + Adjektiv	innere und äussere			3 1,46%	5 0,79%	8 0,56%

[288] Adverbien werden als Adjektive gezählt

Zusammensetzung Untereintrag	Beispiel(e)	Cap	Hausser	Stegmüller	Werlen	Total
Adjektiv + nähere Bestimmung	regulärer, UNIX symmetrische zweier Mengen bewusstes, Sprachlichkeit des abhängig von Sprache	2 0,55%			18 2,86%	20 1,39%
„und“ + Adjektiv	und notwendig			2 0,98%		2 0,14%
Verb (Infinitiv)	lernen				1 0,16%	1 0,07%
„und“ + Verb (Infinitiv)	und verstehen			1 0,49%		1 0,07%
Akronyme/Abkürzungen/ Symbole	VSO $a^k b^k$	3 0,83%	30 12,61%	3 1,46%		36 2,51%
andere	abgeschlossen unter	3 0,83%		7 3,41%	2 0,32%	12 0,84%
Total		361 100%	238 100%	205 100%	630 100%	1434 100%

Tabelle 18: Zusammensetzung der Untereinträge

So wie schon die Vielfalt der Schlagwörter gross war, treten auch die Untereinträge in vielen verschiedenartigen Formen auf. Erneut wurden einmalig erscheinende Untereinträge, die in keine der vorhandenen Kategorien passten unter „andere“ aufgeführt. Die Folgenden wurden erfasst:

abgeschlossen unter
akzeptiert durch Automat
als wissenschaftlich nicht präzisierbar
auf eine Beschreibung relativierte
Bedingung und Ergebnis menschlicher Verständigung
eines Wortes erwerben
grammatische als kognitive Strategien
Intension, Dichotomie, Grundsatz der
semantische Begriffe auf nichtsemantische
und Philosophie der normalen Sprache
Unterschied zu paarweise disjunkt
wissenschaftlich nicht präzisierbar

Bei den Untereinträgen ist auffallend, dass viele elliptisch sind. Die Lücken entstehen dadurch, dass die Schlagwörter nicht wiederholt werden, und können durch die Schlagwörter aufgefüllt werden.

Der Anteil der alleine auftretenden Substantive als Untereinträge beträgt bei allen Registern 25,87%, bei Hausser sogar 47,48%. Damit wird offensichtlich, dass wesentlich mehr allein stehende Substantive auf der Schlagwortebene zu finden sind. Im Gegensatz dazu bestehen 35,22% aller Untereinträge aus einem einzelnen Adjektiv; bei Cap sind es 48,2%. Sie machen somit den grössten Teil der Untereinträge aus, während es bei den Schlagwörtern nur 2,96% einzelne Adjektive waren.

Die Anteile der Verben, Konjunktionen und auch der Rubrik „Akronyme/Abkürzungen/Symbole“ in den Untereinträgen sind verschwindend klein; es dominieren eindeutig die Adjektive und Substantive. 39,82% der Untereinträge besitzen an erster Stelle ein Substantiv, das Zusätze haben kann. Alle Untereinträge, die aus einem einzelnen Adjektiv oder einem Adjektiv und irgendeiner zusätzlichen Angabe gebildet sind, erreichen einen Anteil von 39,54%. Besteht ein Untereintrag aus einem Substantiv und einem Adjektiv, so wird die Voranstellung des Adjektivs innerhalb der Untereinträge vorgezogen.

Präpositionen, die Partikel “als“ sowie die Konjunktion „und“ kommen in 10,33% der Untereinträge bei allen Registern vor, Genitivattribute in 10,17%. Dadurch würde bei einer Ausfilterung aller Funktionswörter mithilfe einer Stoppwortliste sowie einer Lemmatisierung in 20,50% aller Fälle die korrekte Präzisierung der Schlagwörter auf der Ebene der Untereinträge nicht vollständig gelingen.

Der Anteil der Multiwortterme ist bei den Schlagwörtern und bei den Untereinträgen ähnlich gross: bei den Schlagwörtern sind 38,81% aller Schlagwörter Multiwortterme, bei den Untereinträgen sind

es 36,33%. Erneut bescheinigt sich die Verwendung von Phrasen anstelle von einzelnen Wörtern als Indextermkandidaten.

5.2.4 Zusammenfassung

Aus der Untersuchung der fünf Buchregister sowie einer Analyse der Bestandteile der Gesamteinträge (die ich nicht tabellarisch aufgeführt habe) lässt sich schliessen, dass der grösste Anteil von 53,97% aus einem Substantiv als Schlagwort sowie der Anteil von 41,66% der Einträge aus einem Substantiv als Schlagwort und einem Adjektiv als dazugehörigen Untereintrag bestehen. Werden nur die fünf Sachregister ohne die Namenverzeichnisse betrachtet, so beträgt die Portion der einzelnen Substantive als Schlagwort 58,42%, und die aus einem Substantiv als Schlagwort und einem Adjektiv als Untereintrag konstruierten Einträge machen 49,09% aus. Die meisten Tokens in den Registern gehören der Wortklasse der Substantive an, sie werden gefolgt von der Kategorie der Adjektive. Diese Dominanz der Substantive und Adjektive rechtfertigt die Konzentration auf die syntaktischen Nominal- und Präpositionalphrasen, welche auch die Adjektivphrasen subsumieren, bei einer Indexierung mit einem Extraktionsverfahren. Eine Verwendung der lexikalischen Einheiten ist nicht zu empfehlen, da immerhin ein Anteil von 38,81% aller Schlagwörter und ein Anteil von 36,33% aller Untereinträge aus Multiworttermen bestehen.

5.3 Indexierung der Textkörper

In diesem Abschnitt werden die zu den oben untersuchten Buchindexen dazugehörigen Textkörper analysiert und indexiert. Dazu wurde aus jedem der fünf Bücher ein fünfzigseitiger Ausschnitt ausgewählt. Diese Textausschnitte wurden mit dem NP-Chunker von Wojciech verarbeitet und die Nominal- und Präpositionalphrasen sowie deren untergeordnete Nominal- und Präpositionalphrasen unter Beibehaltung der Seitenzahl, auf welchen sie standen, extrahiert. Anschliessend wurden die Funktionswörter entfernt und die verbleibenden Wortgruppen lemmatisiert sowie auf eine kanonische Form gebracht (Kopfsubstantiv an erster Stelle, Adjektive, weitere Substantive usw. nachgestellt). Waren zwei oder mehrere Wortgruppen mit der dazugehörigen Seitenzahl identisch, so wurden sie als eine Wortgruppe angenommen, jedoch die Anzahl der ursprünglich einzelnen Gruppen festgehalten. Die so erzielten Indextermkandidaten wurden dann mit den vorhandenen Buchregistern verglichen, indem für jede extrahierte und normalisierte Art von Wortgruppe die inverse Seitenhäufigkeit berechnet wurde (siehe Abschnitt 5.3.1) und die Kandidaten mit hohen Werten den Indextermen in den Registern gegenübergestellt wurden, um zu prüfen, ob zwischen den hohen Werten der inversen Seitenhäufigkeit und den tatsächlichen Indextermen inklusive Fundstellen in den realen Registern ein Zusammenhang besteht.

Da wir nach der Anwendung der Stoppwortliste und dem Lemmatisierungsprozess nicht mehr von Phrasen sprechen können, werde ich von jetzt an den Begriff *nominale Wortgruppe* verwenden. Zur Veranschaulichung gebe ich hier einen Beispielausschnitt aus der Liste der extrahierten nominalen Wortgruppen mit den dazugehörigen Seitenzahlen von Linke et al. an:

Merkmal 304
Merkmal 341
Merkmal 343
Merkmal 343
Merkmal 343
Merkmal 344
Merkmal 346
Merkmal semantisch 335
Merkmal semantisch 336
Merkmal semantisch 336
Merkmal semantisch 336
Merkmal semantisch 342
Merkmalsbegriff 344
Merkmalsbegriff Semantik 325
Merkmalsbegriff Psycholinguistik 343

In diesem Ausschnitt zählen wir elf verschiedene nominale Wortgruppen. Nur diejenigen, welche dieselben Wörter sowie Seitenzahlen besitzen, werden als eine Wortgruppe zusammengefasst, z.B. „Merkmal semantisch 336“. Die Anzahl der Vorkommen auf einer Buchseite sowie das Total aller Vorkommen auf den verschiedenen Buchseiten werden gespeichert, da sie für die Berechnung der inversen Seitenhäufigkeit gebraucht werden. Wir erhalten in unserem Ausschnitt zwei nominale Wortgruppen, die eine inverse Seitenfrequenz grösser/gleich 1,71 erreichen:

	Vorkommen auf einer Seite	Total Vorkommen	inverse Seitenhäufigkeit
Merkmal 343	3	7	w = 2,56
Merkmal semantisch 336	3	5	w = 3,00

Die fünf Textkörperausschnitte waren alle fünfzig Seiten lang, setzten sich jedoch aus einer unterschiedlichen Anzahl Sätze und Tokens zusammen, wie aus Tabelle 19 ersichtlich ist.

50-seitiger Ausschnitt bei	Anzahl Sätze	Anzahl Tokens
Cap	607	23069
Hausser	741	17972
Linke et al.	807	26039
Stegmüller	237	9368
Werlen	708	20726

Tabelle 19: Anzahl Sätze und Tokens in den Textkörperausschnitten

Jeder der fünf Textausschnitte wurde mit Hilfe der inversen Seitenfrequenz automatisch indexiert und die Indexierungsresultate mit den in den dazugehörigen Registern stehenden Indexterme verglichen, und zwar mit denjenigen Indexeinträgen, deren Fundstellen auf den ausgewählten Textausschnitt verwiesen. Das Ziel der Untersuchung war, herauszufinden, ob sich die inverse Seitenhäufigkeit für das automatische Erstellen von Buchregistern eignet.

Die bestehenden Buchregister werden also als die relevanten Indexterme angesehen. Mit dem Gewichtungsverfahren der inversen Seitenhäufigkeit wird getestet, welche Indexterme gefunden oder welche verpasst werden. Es wird dabei keine Unterscheidung zwischen Schlagwort und Untereintrag gemacht. Sämtliche Untereinträge werden als Schlagwörter behandelt; wo es nötig war, wurden die einschlägigen nicht wiederholten Schlagwörter zu den Untereinträgen dazugeschrieben. Obschon die Einträge der fünf Buchregister als die relevanten Indexterme angenommen werden, so sind diese Register natürlich nicht einfach das Mass aller Dinge. Zumal in allen Registern Fehler gefunden wurden, z.B. Rechtschreibefehler, Formatierungsfehler, Sortierungsfehler.

5.3.1 Die inverse Seitenhäufigkeit

Die inverse Seitenhäufigkeit wurde vom Termgewichtungsverfahren der inversen Dokumenthäufigkeit zur Indexierung von Dokumenten abgeleitet. Sie beinhaltet die statistischen Methoden der Termgewichtung und geht von den folgenden Grundannahmen aus:

- Je häufiger eine nominale Wortgruppe auf einer Buchseite auftritt, desto wichtiger ist sie für das Thema der Seite.
- Je häufiger eine nominale Wortgruppe auf allen Seiten innerhalb des Buches auftritt, desto schlechter diskriminiert sie die Themen der Seiten voneinander.

- Je häufiger eine nominale Wortgruppe auf einer Buchseite auftritt und je weniger sie auf den anderen Seiten des Buches auftritt, desto besser beschreibt sie das Thema der einen Buchseite und ist somit ein guter Indextermkandidat mit einer Referenz auf diese Seite.

Der Gewichtung wurden die linguistischen Methoden der Phrasenextrahierung, Stoppwortausfilterung und Lemmatisierung vorgeschaltet, sodass gesamthaft ein automatisches Extraktionsverfahren mit Termgewichtung zur Erstellung von Schlagwortverzeichnissen für deutschsprachige Texte oder Bücher entstand.

Was bei der inversen Dokumenthäufigkeit ein Term ist, ist bei der inversen Seitenhäufigkeit eine nominale Wortgruppe. Bei der inversen Dokumenthäufigkeit ist ein Dokument das, was bei der inversen Seitenhäufigkeit eine Buchseite ist. Die Kollektion der inversen Dokumenthäufigkeit entspricht dem Total aller Buchseiten der inversen Seitenhäufigkeit. Ein Indexterm der inversen Dokumenthäufigkeit wird bei der inversen Seitenhäufigkeit zum Schlagwort im Register, und der Zeiger zum dazugehörigen Dokument ist bei der inversen Seitenhäufigkeit die Fundstelle mit dem Verweis auf die entsprechende Seite.

Generell stellt die inverse Seitenhäufigkeit die Frequenz einer nominalen Wortgruppe k auf einer Buchseite p zur Anzahl aller Seiten in einem Buch B, in denen die Wortgruppe auftritt, in ein Verhältnis.

$$w_{(k,p)} = \frac{\text{Häufigkeit, mit der k auf einer Seite p auftritt}}{\text{Anzahl der Seiten, in denen k vorkommt}}$$

Das Gewicht eines Terms ist analog zur inversen Dokumenthäufigkeit besonders hoch, wenn es nur wenige Buchseiten gibt, in denen die Wortgruppe auftaucht und wenn sie gleichzeitig auf der entsprechenden Seite oder den entsprechenden Seiten häufiger vorkommt.

Wenn wir nun die kompliziertere Form der inversen Dokumenthäufigkeit, welche die Wortfrequenz und die Dokumentfrequenz berücksichtigt ($tf * idf_{id}$), auf die inverse Seitenhäufigkeit übertragen, so ergibt sich die nachfolgende Formel der inversen Seitenhäufigkeit, welche die Häufigkeit einer nominalen Wortgruppe mit der Seitenfrequenz kombiniert und einer nominalen Wortgruppe auf einer Seite im Verhältnis zur Auftretenshäufigkeit der Wortgruppe auf der Seite und im inversen Verhältnis zur Anzahl der Seiten im Buch, auf welchen die nominale Wortgruppe mindestens einmal auftaucht, ein Gewicht zuweist.

Inverse Seitenhäufigkeit (tf * ipf):

$$w_{kp} = tf_{kp} * \log \left(\frac{N}{n_k} \right)$$

wobei

k = die nominale Wortgruppe
p = die Buchseite
tf_{kp} = Häufigkeit der Wortgruppe k auf einer Seite p
N = Gesamtanzahl aller Seiten in einem Buch B
n_k = Anzahl der Seiten in B, in denen die Wortgruppe k enthalten ist
w_{kp} = Gewicht der inversen Seitenhäufigkeit der nominalen Wortgruppe k auf der Seite p
ipf_k = inverse Seitenfrequenz der Wortgruppe k im Buch B = $\log (N/n_k)$

Das Ganze soll nun wie schon bei der inversen Dokumenthäufigkeit an einem Beispiel verdeutlicht werden. In einem Buch mit 500 Seiten Text beträgt die Anzahl aller Seiten, welche die Wortgruppe k1 enthalten, 25. Auf der Seite p1 tritt k1 9-mal auf, auf einer anderen Seite p2 kommt k1 nur 2-mal vor. Es ergibt sich für k1 in p1 die inverse Seitenhäufigkeit von $w_{k1p1} = 8 * \log (500/25) = 11,71$. In p2 beträgt die inverse Seitenhäufigkeit $w_{k1p2} = 2 * \log (500/25) = 2,60$. Der Wert von k1 in p1 ist um ein Vielfaches grösser als derjenige von k1 in p2.Wenn wir uns vor Augen führen, dass eine Wortgruppe, die in diesem Buch nur ein einziges Mal auf einer einzigen Seite auftritt, eine inverse Seitenhäufigkeit von 2,70 besitzt, wäre k1 in p2 kein Schlagwortkandidat, da dessen Gewicht sogar unter dem Wert von einer einmalig auftretenden Wortgruppe liegt. Die hohe Gewichtung von k1 in p1 macht k1 zu einem Schlagwortkandidaten mit einem Verweis auf die Seite p1.

Bei der Gewichtung von Wortgruppen mit der Methode der inversen Seitenhäufigkeit sollte ein minimaler Grenzwert (w_{min}) festgelegt werden; in unserem Beispiel war das der Wert von 2,70. Das bedeutet, dass eine Wortgruppe, die mehr als nur ein einziges Mal auf einer einzigen Buchseite vorkommt, ein potentielles Schlagwort ist. Da die fünf untersuchten Textausschnitte alle eine Länge von 50 Seiten aufweisen, ergibt sich für alle fünf Bücher der folgende minimale Grenzwert:

$$w_{min} = tf * \log \left(\frac{N}{n} \right) = 1 * \log \left(\frac{50}{1} \right) = 1,70$$

Daraus resultiert eine Festlegung des Grenzwertes auf 1,71, wodurch alle nominalen Wortgruppen, die eine inverse Seitenhäufigkeit von >= 1,71 erreichen, zu Schlagwortkandidaten werden. Jeder Schlagwortkandidat muss somit in einem Textausschnitt gesamthaft mindestens zweimal vorkommen und beide Male auf derselben Seite. Ein maximaler Grenzwert wird nicht definiert, da sich gezeigt hat, dass sich die Resultate mit einer solchen Begrenzung nur verschlechterten. Aber

wir werden im nachfolgenden Abschnitt 5.3.2.3 den minimalen Grenzwert heraufsetzen, um zu ermitteln, ob sich so die besseren Ergebnisse bezüglich des Extrahierens von Schlagwörtern erzielen lassen.

Wir wollen nun zu den erzielten Ergebnissen mit der Gewichtung der nominalen Wortgruppen in unseren fünf Textausschnitten übergehen.

5.3.2 Ergebnisse

In den fünf Textkörperausschnitten wurden die in Tabelle 20 dargestellten nominalen Wortgruppen insgesamt und diejenigen mit einer inversen Seitenhäufigkeit grösser oder gleich dem minimalen Grenzwert von 1,71 gefunden. Dabei ist zu erwähnen, dass identische Wortgruppen, die aus denselben Wörtern mit derselben Seitenangabe bestehen, nur als eine Wortgruppe gezählt werden; zur Berechnung der inversen Seitenhäufigkeit muss jedoch die Anzahl der einzelnen Vorkommen auf jeder Seite sowie des gesamthaften Vorkommens auf allen Buchseiten beibehalten werden. Auch fremdsprachige Nominalgruppen wurden extrahiert und in die Berechnungen mit einbezogen.

Textausschnitt von	Anzahl extrahierte unterschiedliche nominale Wortgruppen	Anzahl nominale Wortgruppen mit tf * ipf >= 1,71	durchschnittlich erzielter tf * ipf-Wert	höchster erzielter tf * ipf-Wert
Cap	2129 100%	262 12,31%	2,88	6,44
Hausser	2915 100%	227 7,79%	2,75	6,29
Linke et al.	4764 100%	259 5,44%	2,91	6,99
Stegmüller	2320 100%	160 6,90%	2,71	5,52
Werlen	3396 100%	179 5,27%	2,69	5,00
Total	15524 100%	1087 7,00%	2,80	6,99

Tabelle 20: Anzahl extrahierte nominale Wortgruppen sowie erzielte tf * ipf-Werte

Der Anteil der extrahierten nominalen Wortgruppen, die eine inverse Seitenhäufigkeit von >= 1,71 aufweisen können, liegt bei allen Büchern innerhalb eines nicht allzu breit gestreuten Bereiches und

beträgt im Schnitt exakt 7%. Auch die Werte der erzielten Höchstgewichte liegen ziemlich nahe beieinander; der minimale Höchstwert beträgt 5,00 bei Werlen und der Maximalwert 6,99 bei Linke et al. Erreicht eine Wortgruppe eine inverse Seitenhäufigkeit von 6,99, so heisst das, dass sie in dem 50-seitigen Ausschnitt total 10-mal auftritt und alle 10 Male auf einer einzigen Seite. Bei einer inversen Seitenhäufigkeit von 5,00 kommt eine nominale Wortgruppe im Textausschnitt gesamthaft 5-mal vor und 5-mal auf derselben Seite. Der Durchschnitt der erzielten Werte der inversen Seitenfrequenz erfährt keine breite Streuung. Bei allen fünf Büchern liegt der Schnitt der Gewichte zwischen minimal 2,69 und maximal 2,91, also innerhalb einer Spannbreite von 0,22.

5.3.2.1 Anpassung der Bewertungsmasse

Bevor wir nun die mit der inversen Seitenhäufigkeit erzielten Schlagwortkandidaten erforschen, werden wir die Bewertungsmasse Recall, Precision, Fallout, Accuracy sowie Error anpassen, um sie für eine Evaluierung der Ergebnisse benutzen zu können. Da wir uns nicht wie im Kapitel „Evaluation eines Index“ mit Dokumenten, sondern mit Indextermen beschäftigen, übertragen wir die Bewertungsmasse auf unsere Untersuchungssituation. Wir halten uns dazu an die Übersicht über die Möglichkeiten innerhalb der Indexterme und der Indextermkandidaten.

	relevante Indexterme	nicht relevante Indexterme	
gefundene Indexterme	a	b	a + b
nicht gefundene Indexterme	c	d	c + d
	a + c	b + d	e

Tabelle 21: Quantitäten der Indexterme in Textausschnitten und dazugehörigen Registern

Die relevanten Indexterme sind die Schlagwörter, die in den Buchregistern stehen und die auf die Seiten der untersuchten Ausschnitte verweisen (alle Untereinträge werden als Schlagwörter betrachtet). Als nicht relevante Indexterme werden diejenigen Indexterme angesehen, welche mit der Methode der inversen Seitenhäufigkeit selektiert wurden, die jedoch nicht in den Buchindexen stehen. Gefundene Indexterme sind die durch die Extraktion und die Anwendung der inversen Seitenfrequenz ermittelten Indextermkandidaten, die nicht gefundenen Indexterme sind die in den Registern auftretenden Schlagwörter, die durch die Extraktion und die Gewichtung nicht ermittelt werden konnten.

Für die Anpassung der Bewertungsmasse sind speziell die folgenden vier Quantitäten von Interesse:

a = Anzahl der Indexterme, die dem Textausschnitt korrekterweise zugeteilt wurden
b = Anzahl der Indexterme, die dem Textausschnitt fälschlicherweise zugeteilt wurden[289]
c = Anzahl der Indexterme, die dem Textausschnitt fälschlicherweise nicht zugeteilt wurden[290]
d = Anzahl der Indexterme, die dem Textausschnitt korrekterweise nicht zugeteilt wurden
e = Anzahl aller aus dem Textausschnitt extrahierten nominalen Wortgruppen

Der **Recall** bzw. die Ausbeute wird nun definiert als das Verhältnis der richtig zugeteilten Indexterme zu der Gesamtanzahl der relevanten Indexterme.

$$R = \frac{a}{a + c}$$

Ein Recall von beispielsweise 0,66 heisst, dass 66% aller relevanten Indexterme richtigerweise gefunden und zugeteilt wurden, dass entsprechend 34% der relevanten Indexterme nicht erwischt wurden.

Das Mass der **Precision** stellt die richtig zugeteilten Indexterme in ein Verhältnis zu der Gesamtanzahl der gefundenen, zugeteilten Indexterme und dient zum Messen der Genauigkeit der Suchmethode und als Indikator für die Fähigkeit des Verfahrens, nicht relevante Indexterme auszuscheiden.

$$P = \frac{a}{a + b}$$

Ein Precisionwert von 0,52 sagt aus, dass 52% aller gefundenen zugeteilten Indexterme auch relevant waren.

Der **Fallout** stellt das Mass für den Ballast dar und ist definiert als das Verhältnis der zugeteilten nicht relevanten Indexterme zur Gesamtzahl aller nicht relevanten Indexterme im Buch.

[289] b bezieht sich auf Kommissionsfehler
[290] c bezieht sich auf Unterlassungsfehler

$$F = \frac{b}{b + d}$$

Von allen nicht relevanten Indextermen beträgt der Anteil der fälschlicherweise gefundenen Indexterme bei einem Fallout von 0,14 also 14%. Der Wert sollte möglichst klein sein.

Das Mass der **Accuracy** bezieht sich auf die Treffgenauigkeit der gefundenen nominalen Wortgruppen. Die Anzahl aller gefundenen relevanten oder korrekterweise nicht gefundenen nicht relevanten Indexterme wird in ein Verhältnis zu der Gesamtanzahl aller nominalen Wortgruppen in einem Buchausschnitt gesetzt.

$$A = 1 - \text{Error} = \frac{a + d}{a + b + c + d}$$

Je höher der Wert, desto besser die Effektivität der Methode. Ein Accuracywert von 0,91 bedeutet in unserem Fall, dass bei einer automatischen Indexierung mit der Gewichtungsmethode der inversen Seitenhäufigkeit 91% Genauigkeit erreicht wurde oder dass 91% der relevanten Indexterme gefunden sowie der nicht relevanten Indexterme korrekterweise nicht zugeteilt wurden.

Das Bewertungsmass **Error** bezieht sich sowohl auf Unterlassungs- wie auch Kommissionsfehler, sein Wert sollte möglichst klein sein.

$$E = 1 - \text{Accuracy} = \frac{b + c}{a + b + c + d}$$

Die Fehlerrate definiert sich als das Verhältnis aller gefundenen nicht relevanten plus aller nicht gefundenen relevanten Indexterme zur Anzahl sämtlicher nominaler Wortgruppen in einem Textausschnitt. Ein Error von 0,09 heisst, dass 9% aller Indexterme bei einer automatischen Indexierung fälschlicherweise zugeteilt oder zurückbehalten wurden. Accuracy und Error ergeben zusammen immer 1 bzw. 100%, wie wir in den Abschnitten 4.4.6 und 4.4.7 schon gesehen haben.

5.3.2.2 Die einzelnen Textausschnitte

Im Folgenden wird für jedes untersuchte und indexierte Buch eine Übersicht über die Verteilung der relevanten und nicht relevanten sowie der gefundenen und nicht gefundenen Indexterme plus die erzielten Bewertungsmasswerte angegeben.

Zur Auszählung der Treffer wurde ein Indexterm als gefunden und relevant beurteilt, wenn eine nominale Wortgruppe aus einem Textausschnitt mit dem Eintrag in dem dazugehörigen Register nicht nur bezüglich der Wörter, sondern auch bezüglich der Fundstelle vollständig übereinstimmte, jedoch ohne Berücksichtigung der Funktionswörter, da diese ja aus den extrahierten Schlagwortkandidaten entfernt wurden. Als Treffer gelten ebenso diejenigen nominalen Wortgruppen, welche in Bezug auf das Ordnungswort und die Fundstelle mit einem Eintrag im Register übereinstimmen, die nähere Bestimmung zum Ordnungswort jedoch geringfügig abweicht. Das heisst, dass beispielsweise eine nähere Bestimmung im einen Fall aus dem Adjektiv („semantisch“), im anderen Fall aber dem Substantiv desselben Lexems bestand („Semantik“). Als Treffer nicht gezählt wurden extrahierte Wortgruppen, die mit den Termen eines Eintrags identisch waren, deren Seitenzahlen bzw. Fundstellen indessen voneinander abwichen. Für die Ermittlung der Treffer wurden alle Schlagwörter in den fünf Buchregistern, die aus einem vorangestellten Adjektiv (oder mehreren) und einem Substantiv gebildet waren, auf die invertierte Form mit nachgestelltem Adjektiv gebracht. Wie schon oben erwähnt wurden nur diejenigen Wortgruppen mit einer inversen Seitenhäufigkeit von >= 1,71 berücksichtigt.

Tabelle 22 zeigt die Verhältnisse der Indexterme und Indextermkandidaten beim Textausschnitt von Cap auf.

	relevante Indexterme	nicht relevante Indexterme	
gefundene Indexterme	a 44	b 218	a + b 262
nicht gefundene Indexterme	c 88	d 1779	c + d 1867
	a + c 132	b + d 1997	e 2129

Tabelle 22: Quantitäten der Indexterme bei Cap (w_{min} >= 1,71)

Mit diesen Quantitäten wurden die Werte unserer fünf Evaluationsmasse der folgenden Tabelle 23 bei einem minimalen Grenzwert von 1,71 erzielt.

$\text{Recall}_{\text{Cap}}$	=	$\frac{a}{a + c}$	=	0,33
$\text{Precision}_{\text{Cap}}$	=	$\frac{a}{a + b}$	=	0,17
$\text{Fallout}_{\text{Cap}}$	=	$\frac{b}{b + d}$	=	0,11
$\text{Accuracy}_{\text{Cap}}$	=	$\frac{a + d}{a + b + c + d}$	=	0,86
$\text{Error}_{\text{Cap}}$	=	$\frac{b + c}{a + b + c + d}$	=	0,14

Tabelle 23: Erreichte Evaluationswerte bei Cap

Der erreichte Recallwert von 0,33, der besagt, dass nur gerade 33% aller relevanten Indexterme richtig zugeteilt wurden, ist zwar klein, aber immerhin höher als der erzielte Precisionwert von 0,17, wonach nur gerade 17% aller gefundenen Indexterme auch relevant waren. Entsprechend hoch ist auch der Fallout mit 0,11; von allen nicht relevanten Indextermen wurden 11% fälschlicherweise als Ballast zugeteilt. Erstaunlich hoch ist die Treffgenauigkeit der Zuteilungen mit einem Accuracywert von 0,86 und einer Errorrate von 0,14. Das bedeutet, dass die Genauigkeit der zugeteilten oder nicht zugeteilten Indexterme recht hoch liegt bzw. dass die Fehlerrate 14% beträgt. Obschon die Treffgenauigkeit einen relativ hohen Wert erreicht, bleibt zu bemerken, dass der Recall und vor allem die Precision unbefriedigend sind.

Die beim Textausschnitt von Hausser erzielten Quantitäten der Indexterme sind in Tabelle 24 dargestellt.

	relevante Indexterme	nicht relevante Indexterme	
gefundene Indexterme	a 72	b 155	a + b 227
nicht gefundene Indexterme	c 146	d 2542	c + d 2688
	a + c 218	b + d 2697	e 2915

Tabelle 24: Quantitäten der Indexterme bei Hausser (w_{min} >= 1,71)

Daraus lassen sich die Werte der Evaluationsmasse in Tabelle 25 errechnen.

Recall $_{Hausser}$	=	$\frac{a}{a + c}$	= 0,33
Precision $_{Hausser}$	=	$\frac{a}{a + b}$	= 0,32
Fallout $_{Hausser}$	=	$\frac{b}{b + d}$	= 0,06
Accuracy $_{Hausser}$	=	$\frac{a + d}{a + b + c + d}$	= 0,90
Error $_{Hausser}$	=	$\frac{b + c}{a + b + c + d}$	= 0,10

Tabelle 25: Erreichte Evaluationswerte bei Hausser

Mit dem Termgewichtungsverfahren der inversen Seitenfrequenz erreicht man im Textausschnitt von Hausser ebenso wie bei Cap einen Recall von nur 0,33. Im Gegensatz zu Cap beträgt die Präzision beim Ausschnitt von Hausser jedoch 0,32, also fast das Doppelte. 32% aller gefundenen Indextermkandidaten waren also relevant. Der Fallout oder Ballast beträgt bei Hausser mit 0,06 bedeutend weniger als bei Cap. Der Wert der Accuracy ist geringfügig höher (Cap 0,86; Werlen 0,90), der Error entsprechend etwas niedriger (Cap 0,14; Werlen 0,10). Auch wenn die Werte der

Evaluationsmasse bessere Resultate bei Hausser erzielen, so ist ein Recall von 0,33 und eine Precision von 0,32 noch nicht zufrieden stellend.

Tabelle 26 stellt die gefundenen und nicht gefundenen, die relevanten und nicht relevanten Indexterme beim Ausschnitt von Linke et al. dar, die Tabelle 27 die dazugehörigen Werte der Performanzbewertung.

	relevante Indexterme	nicht relevante Indexterme	
gefundene Indexterme	a 48	b 211	a + b 259
nicht gefundene Indexterme	c 52	d 4453	c + d 4505
	a + c 100	b + d 4664	e 4764

Tabelle 26: Quantitäten der Indexterme bei Linke et al. (w_{min} >= 1,71)

Recall $_{\text{Linke et al.}}$	=	$\frac{a}{a + c}$	= 0,48
Precision $_{\text{Linke et al.}}$	=	$\frac{a}{a + b}$	= 0,19
Fallout $_{\text{Linke et al.}}$	=	$\frac{b}{b + d}$	= 0,08
Accuracy $_{\text{Linke et al.}}$	=	$\frac{a + d}{a + b + c + d}$	= 0,94
Error $_{\text{Linke et al.}}$	=	$\frac{b + c}{a + b + c + d}$	= 0,06

Tabelle 27: Erreichte Evaluationswerte bei Linke et al.

Mit dem Buchausschnitt von Linke et al. gelingt ein Recall von 0,48. Der Wert ist um 0,15 höher als bei Cap und Hausser, er kommt an die 50%-Marke jedoch nicht heran. Die Präzision der gefundenen Indexterme beträgt enttäuschende 0,19 und liegt damit im Bereich von Cap. Der

Falloutwert befindet sich mit 0,08 etwa in der Mitte zwischen den Ballastanteilen bei Cap und Hausser. Diejenigen Indexterme, die mit der inversen Seitenfrequenz ermittelt werden können, werden mit einer hohen Treffgenauigkeit von 0,94 zugewiesen. Nur in 6% aller Fälle werden die gefundenen Wortgruppen fehlerhaft zugeteilt oder nicht zugeteilt. Die Accuracy- und Errorwerte können allerdings den nicht zufrieden stellenden Recall keineswegs wettmachen.

Im Textkörperausschnitt von Stegmüller sehen die Ergebnisse folgendermassen aus:

	relevante Indexterme	nicht relevante Indexterme	
gefundene Indexterme	a 42	b 118	a + b 160
nicht gefundene Indexterme	c 102	d 2058	c + d 2160
	a + c 144	b + d 2176	e 2320

Tabelle 28: Quantitäten der Indexterme bei Stegmüller (w_{min} >= 1,71)

Mit diesen Indextermen und Indextermkandidaten wurden die in der Tabelle 29 aufgelisteten Effizienzmesswerte erzielt. Der minimale Grenzwert ist immer noch auf >= 1,71 festgelegt.

Recall $_{Stegmüller}$	=	$\frac{a}{a + c}$	=	0,29
Precision $_{Stegmüller}$	=	$\frac{a}{a + b}$	=	0,26
Fallout $_{Stegmüller}$	=	$\frac{b}{b + d}$	=	0,05
Accuracy $_{Stegmüller}$	=	$\frac{a + d}{a + b + c + d}$	=	0,91
Error $_{Stegmüller}$	=	$\frac{b + c}{a + b + c + d}$	=	0,09

Tabelle 29: Erreichte Evaluationswerte bei Stegmüller

Auch beim Text von Stegmüller kann die inverse Seitenhäufigkeit weder die Ausbeute noch die Präzision zu einem erfolgreichen Resultat bringen. Nur 29% aller relevanten Einträge können ermittelt werden, ausschliesslich 26% aller gefundenen nominalen Wortgruppen sind relevant. Der Fallout befindet sich mit 0,05 in einem akzeptablen Bereich. Abermals stehen die Treffgenauigkeit und die Fehlerrate in einem recht guten Verhältnis.

Die Tabelle 30 zeigt nun noch die Quantitäten der Indexterme beim fünften Textausschnitt auf, demjenigen von Werlen.

	relevante Indexterme	nicht relevante Indexterme	
gefundene Indexterme	a 80	b 99	a + b 179
nicht gefundene Indexterme	c 438	d 2779	c + d 3217
	a + c 518	b + d 2878	e 3396

Tabelle 30: Quantitäten der Indexterme bei Werlen (w_{min} >= 1,71)

In der Tabelle 31 stehen die dazugehörigen erzielten Effizienzmasswerte.

Recall $_{Werlen}$	=	$\frac{a}{a + c}$	= 0,15
Precision $_{Werlen}$	=	$\frac{a}{a + b}$	= 0,45
Fallout $_{Werlen}$	=	$\frac{b}{b + d}$	= 0,03
Accuracy $_{Werlen}$	=	$\frac{a + d}{a + b + c + d}$	= 0,84
Error $_{Werlen}$	=	$\frac{b + c}{a + b + c + d}$	= 0,16

Tabelle 31: Erreichte Evaluationswerte bei Werlen

Während die Precision bei Werlen den bisher höchsten Wert von 0,45 erreicht, beträgt der Recall nur 0,15. Dies ist die niedrigste Ausbeute von allen fünf Textausschnitten. Auch der gelieferte Ballast von nur 3% aller nicht einschlägigen Wortgruppen kann über den schlechten Recall nicht hinwegtäuschen. Die Treffgenauigkeit von 0,84 ist den anderen vier Büchern unterlegen, parallel dazu fällt die Fehlerrate von 0,16 bei Werlen am meisten ins Gewicht; sie wird hauptsächlich durch die schlechte Ausbeute ausgelöst.

Die ausgeglichensten Ergebnisse bezüglich Recall und Precision werden mit der inversen Seitenfrequenz im Textausschnitt von Hausser erreicht. Sowohl die Ausbeute wie auch die Präzision erreichen in diesem Buch beide über 30%. Den extremsten Unterschied zwischen den beiden Massen findet sich bei Werlen. Seiner ist auch der einzige Textausschnitt, bei welchem der Precisionwert höher liegt als der Recallwert. Der Bereich des Fallout bewegt sich zwischen 0,03 bei Werlen und 0,11 bei Cap; die anderen Werte liegen dazwischen. Das Buch von Cap ist meiner Ansicht nach das einzige, welches eine akzeptable Menge an geliefertem Ballast überschreitet. Die 8% von Linke et al., die 6% von Hausser sowie die 5% von Stegmüller sind gerade noch in Ordnung, der Anteil von 3% Ballast bei Werlen ist erstrebenswert. Die Proportionen von Accuracy und Error verhalten sich bei den fünf Büchern ähnlich. Die Treffgenauigkeit liegt bei allen recht hoch (über 0,84), das beste Ergebnis erzielen Linke et al. mit 0,94 Accuracy und 0,06 Error.

Nach der Indexierung der einzelnen Buchausschnitte mit der Gewichtungsmethode der inversen Seitenhäufigkeit zeichnet sich der Schluss ab, dass die alleinige Anwendung der inversen Seitenhäufigkeit mit den einfachen linguistischen Vorverarbeitungsprozessen für die automatische Indexierung von deutschsprachigen Texten keine zufrieden stellenden Resultate in Bezug auf die Effizienz liefert und die Methode somit ohne zusätzliche Verfahren nicht geeignet zu sein scheint.

5.3.2.3 Alle Textausschnitte

Wir betrachten im Folgenden die Ergebnisse aller fünf Textausschnitte zusammen. Tabelle 32 stellt die Mengenverhältnisse der Indexterme und Indextermkandidaten von allen analysierten Büchern dar.

	relevante Indexterme	nicht relevante Indexterme	
gefundene Indexterme	a 286	b 801	a + b 1087
nicht gefundene Indexterme	c 826	d 13611	c + d 14437
	a + c 1112	b + d 14412	e 15524

Tabelle 32: Quantitäten der Indexterme bei allen Textausschnitten (w_{min} >= 1,71)

Es wurden gesamthaft die folgenden Werte unserer fünf Evaluationsmasse erzielt.

Recall $_{gesamthaft}$	=	$\frac{a}{a + c}$	= 0,26
Precision $_{gesamthaft}$	=	$\frac{a}{a + b}$	= 0,26
Fallout $_{gesamthaft}$	=	$\frac{b}{b + d}$	= 0,06
Accuracy $_{gesamthaft}$	=	$\frac{a + d}{a + b + c + d}$	= 0,90
Error $_{gesamthaft}$	=	$\frac{b + c}{a + b + c + d}$	= 0,10

Tabelle 33: Erreichte Evaluationswerte bei allen Textausschnitten (w_{min} >= 1,71)

Es ist offensichtlich, dass alle indexierten Textausschnitte zusammen ein ausgeglichenes Verhältnis von Recall und Precision erreichen: 26% aller relevanten Indexterme werden durch die inverse Seitenhäufigkeit gefunden, und 26% aller gefundenen Indexterme sind relevant. Auch wenn das Verhältnis der beiden Masse ausbalanciert ist, so bilden die Werte kein befriedigendes Ergebnis. Nicht einmal ein Drittel aller relevanten Indexterme wurde ermittelt, und weniger als ein Drittel aller gefundenen Indexterme ist tatsächlich einschlägig. Der Anteil an geliefertem Ballast von 6% erscheint mir eher unproblematisch. Auch eine Treffgenauigkeit von 0,90 sowie eine Fehlerrate von

0,10 sind annehmbar. Das Hauptproblem der Indexierung mit der inversen Seitenhäufigkeit ist eindeutig die Ausbeute. Sie steht in dieser Untersuchung im Gegensatz z.B. zur Arbeit von Lahtinen im Vordergrund. Lahtinen versucht in seinem Ansatz des automatischen Indexierens, welcher auch linguistische und statistische Methoden kombiniert, einen möglichst hohen Precisionwert anzustreben, weil nur die besten Deskriptoren identifiziert werden sollen.[291] Beim Indexieren mit der inversen Seitenhäufigkeit ist es das Ziel, einen besseren Recall zu erreichen, da es zuerst einmal gilt, möglichst viele Indextermkandidaten zu ermitteln, um erst in einem zweiten Schritt die tatsächlich relevanten zu identifizieren. Ausserdem fallen Ballast und Treffgenauigkeit angemessen aus.

Bisher wurde der minimale Grenzwert auf 1,71 festgelegt. Wird dieser höher angesetzt, so sollten sich die Präzision verbessern und der Ballastanteil verringern lassen. Die Tabellen 34 und 35 zeigen die Mengenverhältnisse und Effizienzwerte auf, die mit einem minimalen Grenzwert von 2,01 erzielt wurden. Als Indextermkandidaten wurden also nur nominale Wortgruppen berücksichtigt, welche eine inverse Seitenhäufigkeit von grösser/gleich 2,01 besassen. Sämtliche Textausschnitte werden gemeinsam erfasst.

	relevante Indexterme	nicht relevante Indexterme	
gefundene Indexterme	a 241	b 711	a + b 952
nicht gefundene Indexterme	c 871	d 13701	c + d 14574
	a + c 1112	b + d 14412	e 15524

Tabelle 34: Quantitäten der Indexterme bei allen Textausschnitten (w_{min} >= 2,01)

[291] Lahtinen 2000:178

Recall $_{gesamthaft}$	=	$\frac{a}{a + c}$	=	0,22
Precision $_{gesamthaft}$	=	$\frac{a}{a + b}$	=	0,25
Fallout $_{gesamthaft}$	=	$\frac{b}{b + d}$	=	0,05
Accuracy $_{gesssamthaft}$	=	$\frac{a + d}{a + b + c + d}$	=	0,90
Error $_{gesamthaft}$	=	$\frac{b + c}{a + b + c + d}$	=	0,10

Tabelle 35: Erreichte Evaluationswerte bei allen Textausschnitten (w_{min} >= 2,01)

Sowohl Recall wie auch Precision verschlechtern sich mit dem höher angesetzten minimalen Grenzwert, die Precision allerdings sehr gering. Der Fallout oder Ballastanteil wird unwesentlich kleiner. Accuracy und Error bleiben gleich. Ich ziehe daraus den Schluss, dass es nicht sinnvoll und ergebnisverbessernd ist, den minimalen Grenzwert höher als bei 1,71 festzulegen. Wir erinnern uns, dass dieser Wert jede nominale Wortgruppe als Indextermkandidat oder gefundenen Indexterm betrachtet, welche im gesamten Textausschnitt zweimal, aber beide Male nur auf dieser einen Seite vorkommen. Die Resultate, vor allem der Recall, verschlechtern sich bei einem minimalen Grenzwert von 2,01 oder höher. Doch genau die Verbesserung des Recallwertes wünsche ich anzustreben. Auf die Darstellung der Ergebnisse mit noch höheren minimalen Grenzwerten werde ich verzichten, sie trugen zu einer weiteren Verschlechterung sämtlicher Evaluationsmasse bei.

In den folgenden drei Abschnitten werden wir nun einen Bezug zwischen der Indexierung der Textausschnitte mithilfe der inversen Seitenhäufigkeit und der Untersuchung der Bestandteile in den Buchregistern im Abschnitt 5.2 herstellen. Wir wollen die gefundenen relevanten Indexterme, die nicht gefundenen, aber relevanten Indexterme und die gefundenen nicht relevanten Indexterme hinsichtlich ihrer lexikalischen und syntaktischen Bestandteile analysieren, um Erkenntnisse bezüglich einer möglichen Modifizierung unseres Indexierungsverfahrens mit dem Ziel einer Performanzverbesserung zu gewinnen.

5.3.2.3.1 Analyse der gefundenen relevanten Indexterme

Bevor wir uns mit den lexikalischen Wortkategorien oder den syntaktischen Konstituenten der ermittelten einschlägigen Indexterme beschäftigen, möchte ich kurz einen Überblick über die mit dem Verfahren der inversen Seitenhäufigkeit erzielten Gewichte der gefundenen relevanten Indexterme geben und untersuchen, in welchem Verhältnis diese zu den extrahierten nominalen Wortgruppen stehen. Tabelle 36 enthält die dazu notwendigen Informationen.

Textausschnitt von	Anzahl extrahierte unterschiedliche nominale Wortgruppen	Anzahl gefundene relevante Indexterme mit tf * ipf >= 1,71 (Treffer)	durchschnittlich erzielter tf * ipf-Wert bei den Treffern	höchster erzielter tf * ipf-Wert bei den Treffern
Cap	2129 100%	44 2,07%	2,73	4,00
Hausser	2915 100%	72 2,47%	2,77	4,60
Linke et al.	4764 100%	48 1,01%	3,16	5,62
Stegmüller	2320 100%	42 1,81%	2,75	5,52
Werlen	3396 100%	80 2,36%	2,77	4,89
allen	15524 100%	286 1,84%	2,84	5,62

Tabelle 36: Anzahl der gefundenen relevanten Indexterme und erzielte tf * ipf-Werte

Von den insgesamt extrahierten 15'524 nominalen Wortgruppen sind nur 1,84% oder 286 relevante Indexterme. Der durchschnittliche Wert der inversen Seitenhäufigkeit der gefundenen relevanten Terme beträgt 2,84. Eine Nominalgruppe mit einem Gewicht von 2,84 kommt in dem fünfzigseitigen Textausschnitt in etwa zehnmal insgesamt vor und auf der Seite mit dem erreichten Gewicht viermal. Der durchschnittliche tf*ipf-Wert von allen extrahierten Nominalgruppen (>= 1,71) beträgt 2,80. Es lässt sich kein markanter Unterschied zwischen dem durchschnittlichen Gewicht aller Wortgruppen und dem durchschnittlichen Gewicht der Treffer konstatieren. Die im Schnitt ermittelten Gewichte bei den einzelnen Büchern liegen ziemlich nahe beieinander und bewegen sich zwischen 2,73 und 3,16. Ebenso betragen die durchschnittlichen Gewichte aller Wortgruppen minimal 2,69 und maximal 2,91. Der höchste Wert eines relevanten gefundenen Indexterms beträgt 5,62. Der kleinste errechnete Wert entspricht natürlich dem minimalen Grenzwert von 1,71. Der höchste erzielte Wert aller extrahierten Wortgruppen beläuft sich auf 6,99

und ist damit grösser als der höchste Wert eines Treffers. Eine Verbesserung der Präzision lässt sich durch die Festsetzung eines maximalen Grenzwertes nicht erreichen, denn das höchste erreichte durchschnittliche Gewicht bei Cap betrifft die inverse Seitenhäufigkeit einer relevanten Wortgruppe und ebenso diejenige einer gefundenen, aber nicht relevanten Wortgruppe. Ausserdem verschlechterten sich die Ergebnisse gesamthaft durch die Einführung eines maximalen Grenzwertes (wie schon im letzten Abschnitt beschrieben).

Die lexikalischen und syntaktischen Bestandteile der gefundenen relevanten Indexterme sind in Tabelle 37 aufgelistet. Ein „S“ bedeutet Substantiv, „A“ ist die Abkürzung für Adjektiv, „G“ steht für Genitivattribut, „P“ für Präpositionalphrase, „Ak/Ab/Sy“ heisst Akronym, Abkürzung und Symbol.

Wort-kategorien	Cap	Hausser	Linke et al.	Stegmüller	Werlen	alle
S	25 56,82%	40 55,55%	36 75,00%	14 33,33%	60 75,00%	175 61,19%
S + S	7 15,91%	11 15,28%	-	15 35,72%	9 11,25%	42 14,68%
S + S + S	2 4,55%	-	-	-	-	2 0,70%
S + A	10 22,72%	19 26,39%	12 25,00%	10 23,81%	9 11,25%	60 20,98%
S + 2 A	-	1 1,39%	-	1 2,38%	1 1,25%	3 1,05%
S + G	-	-	-	1 2,38%	1 1,25%	2 0,70%
S + P	-	-	-	1 2,38%	-	1 0,35%
Ak/Ab/Sy	-	1 1,39%	-	-	-	1 0,35%
Total	44 100%	72 100%	48 100%	42 100%	80 100%	286 100%

Tabelle 37: Bestandteile der gefundenen relevanten Indexterme

Die relevanten gefundenen Indexterme oder die Treffer bestehen gesamthaft zu 61,19% aus einem einzigen Substantiv, in zwei Büchern sogar zu 75%. Einen Anteil von 61,54% aller Treffer bilden einzelne Wörter. Ein einzelnes Adjektiv konnte nicht gefunden werden, da die Adjektivphrasen der Textausschnitte nicht isoliert als Indextermkandidaten ermittelt wurden, sondern immer nur als

untergeordnete Phrasen zu einer Nominal- oder Präpositionalphrase. Somit konnten 2,61% aller einschlägigen Schlagwörter, die sich aus einem oder zwei Adjektiven zusammensetzen, nicht gefunden werden. Einzelne Verben würden mit der Vorgehensweise ebenso nicht ermittelt, bei den relevanten Schlagwörtern gab es jedoch keine Verben.

38,46% aller Treffer sind Mehrwortterme. Da es sich bei den Kombinationen um Substantiv plus ein oder zwei Substantive, Substantiv plus ein oder mehrere Adjektive sowie Substantiv plus Genitivattribut oder Präpositionalphrase handelt, konnten sie durch die Verwendung des NP-Chunkers gefunden werden. Auch die Trefferanalyse bestätigt die Verwendung von Phrasen anstelle von einzelnen Wörtern als Indextermkandidaten. Obwohl die Genitivattribute und die untergeordneten Präpositonalphrasen nur einen äussert geringen Anteil aller relevanten Indextermbestandteile ausmachen, so sind doch die attributiv verwendeten Adjektive zur Erfassung der Registereinträge bedeutend sowie auch die Treffer, die aus zwei oder mehreren Substantiven gebildet sind. Ausserdem können Multiwortterme nach Auffassung von Lahtinen gebraucht werden, um die Präzision eines Retrievals zu verbessern, weil Phrasen spezifischer sind als einzelne Wörter.[292] Ich habe allerdings nicht geprüft, wie die Ergebnisse durch die Verwendung von einzelnen Wörter im Gegensatz zur Verwendung von Wortgruppen ausgesehen hätten.

Der Anteil der korrekterweise gefundenen und zugeteilten Indexterme im Verhältnis zu allen gefundenen Wortgruppen bzw. die Precision beträgt bei allen Büchern gesamthaft nur 26,31%.

5.3.2.3.2 Analyse der nicht gefundenen relevanten Indexterme

In diesem Abschnitt werden die mit unserem Verfahren nicht ermittelten Indexterme, die jedoch relevant wären, diskutiert. Wir schauen uns sozusagen den Informationsverlust genauer an. Die Tabelle 38 gibt Auskunft über die Zusammensetzung der verpassten Schlagwortregistereinträge. Die verwendeten Abkürzungen sind analog zur Tabelle im vorangehenden Abschnitt. Zusätzlich bedeutet „n.B.“ nähere Bestimmung, eckige Klammern drücken Zusammengehörigkeit aus.

[292] Lahtinen 2000:177

Wort-kategorien	Cap	Hausser	Linke et al.	Stegmüller	Werlen	alle
S	31 35,23%	40 27,40%	39 75,00%	14 13,73%	196 44,75%	320 38,74%
S + S (als n.B.)	13 14,77%	33 22,60%	-	26 25,49%	52 11,87%	124 15,01%
S + [S + S als n.B.]	4 4,55%	4 2,74%	-	-	4 0,91%	12 1,45%
S + A	17 19,32%	43 29,45%	12 23,08%	30 29,41%	88 20,09%	190 23,00%
S + 2 A	1 1,14%	-	1 1,92%	5 4,90%	5 1,14%	12 1,45%
S + A + [S als n.B.]	-	3 2,05%	-	3 2,94%	2 0,46%	8 0,97%
S + [A + S als n.B.]	-	3 2,05%	-	-	24 5,48%	27 3,27%
S + Gen	2 2,27%	3 2,05%	-	10 9,80%	43 9,82%	58 7,02%
S + PP	2 2,27%	2 1,37%	-	6 5,88%	21 4,79%	31 3,75%
S + [Ak/Ab/Sy als n.B.]	-	6 4,11%	-	-	-	6 0,73%
A	9 10,23%	3 2,05%	-	4 3,92%	3 0,68%	19 2,30%
A + A	-	6 4,11%	-	4 3,92%	-	10 1,21%
A + [S als n.B.]	5 5,68%	-	-	-	-	5 0,61%
Ak/Ab/Sy	4 4,55%	-	-	-	-	4 0,48%
Total	88 100%	146 100%	52 100%	102 100%	438 100%	826 100%

Tabelle 38: Bestandteile der relevanten nicht gefundenen Indexterme

Der Anteil der nicht gefundenen relevanten Indexterme, die aus einem einzelnen Substantiv bestehen, beträgt 38,74%. Wenn wir bedenken, dass 44,51% aller relevanten Schlagwörter aus einem einzelnen Substantiv bestehen, so ist der hohe Anteil der verpassten Substantive nicht

verwunderlich. Ebenso verhält es sich mit den Indextermen, die sich aus einem Substantiv und einem Adjektiv zusammensetzen. 22,48% der relevanten Schlagwörter sind aus einem Substantiv und einem Adjektiv gebildet, wovon ein Anteil von 76% nicht gefunden werden konnte respektive 23% aller nicht gefundenen, aber relevanten Indexeinträge aus der Zusammensetzung Substantiv plus Adjektiv bestehen. Die Indexeinträge Substantiv plus Genitivattribut oder Substantiv plus Präpositionalphrase können im Vergleich zu den Treffern (1,05% Anteil) häufiger nicht gefunden als identifiziert werden. Diese Klasse beläuft sich auf einen Anteil von 10,77% der verpassten Einträge. 95,88% aller nicht ermittelten einschlägigen Indexterme besitzen als Ordnungswort ein Substantiv. Die Tatsache, dass sie verpasst wurden, hat nichts mit dem Vorgehen der Nominal- und Präpositionalphrasenextraktion zu tun, der Grund für den Informationsverlust muss anderswo liegen, z.B. bei einem nicht genügend häufigen Auftreten auf einer Seite des Textausschnittes oder bei Formulierungsvariationen zur Vermeidung von Wiederholungen im Text. Die verbleibenden 4,12% der verpassten Einträge sind einzelne Adjektive, zwei Adjektive oder ein Adjektiv plus ein Substantiv. Sie konnten durch den Entscheid für eine Extraktion der Nominal- und Präpositionalphrasen unmöglich gefunden werden. Eine Verwendung aller Adjektivphrasen als Indextermkandidaten erscheint mir ungünstig; der Ballast dürfte erheblich grösser sein als der dadurch verhinderte Informationsverlust.

Betrachten wir den Anteil der einzelnen Wörter sowie der Mehrwortterme beim Informationsverlust, so ergeben sich die folgenden Verhältnisse: 41,53% einzelne Wörter, 58,47% Mehrwortterme. Die Portion der nicht gefundenen relevanten Mehrwortterme ist somit grösser als der Anteil bei den Treffern der Mehrwortterme (38,46%). Der Umstand hat seine Ursache darin, dass mehrgliedrige Wortgruppen auf vielfältigere Weise formuliert werden können als einzelne Wörter oder Fachtermini und deshalb mit der Termhäufigkeitsmethode nicht gefunden werden konnten. Um den Anteil der Treffer in Bezug auf die Mehrwortterme und auch auf die einzelnen Wörter verbessern zu können, müsste die Termgewichtungsmethode der inversen Seitenfrequenz die Mannigfaltigkeit der Ausdrucksweisen berücksichtigen, z.B. durch die Verwendung eines Thesaurus oder die Konzentration auf die Kopfwörter der extrahierten Phrasen bei der Berechnung der inversen Seitenhäufigkeit.

In allen fünf Textausschnitten zusammen konnten 74,28% aller relevanten Indexterme mit dem Gewichtungsverfahren der inversen Seitenhäufigkeit nicht gefunden werden (c dividiert durch a+c). Dies ist eine sehr hohe Informationsverlustquote; das Ergebnis ist keineswegs befriedigend.

5.3.2.3.3 Analyse der gefundenen nicht relevanten Indexterme

In diesem Abschnitt wollen wir nun auch noch die Indexeinträge untersuchen, die fälschlicherweise zugeteilt wurden. Wir analysieren also die gefundenen, jedoch nicht relevanten Indexterme oder den

Ballast. In der Tabelle 39 sind abermals die lexikalischen und syntaktischen Bestandteile der betreffenden Indexterme dargestellt.

Wort-kategorien	Cap	Hausser	Linke et al.	Stegmüller	Werlen	alle
S	114 52,29%	84 54,19%	161 76,30%	76 64,41%	64 64,65%	499 62,30%
S + S	46 21,10%	8 5,16%	21 9,95%	11 9,32%	11 11,11%	97 12,11%
S + S + S	1 0,46%	2 1,29%	3 1,42%	-	1 1,01%	7 0,87%
S + A	40 18,35%	38 24,52%	22 10,43%	18 15,25%	18 18,18%	136 16,98%
S + 2 A	4 1,83%	3 1,94%	2 0,95%	4 3,39%	1 1,01%	14 1,75%
[S + A] + S	5 2,29%	1 0,65%	-	3 2,54%	1 1,01%	10 1,25%
S + [A + S]	6 2,75%	10 6,45%	2 0,95%	5 4,24%	1 1,01%	24 3,00%
[S + A] + [S + A]	-	4 2,58%	-	1 0,85%	2 2,02%	7 0,87%
Ak/Ab/Sy	2 0,92%	5 3,23%	-	-	-	7 0,87%
Total	218 100%	155 100%	211 100%	118 100%	99 100%	801 100%

Tabelle 39: Bestandteile der gefundenen nicht relevanten Indexterme

Den grössten Anteil der fälschlicherweise zugeteilten Indexterme machen erneut die einzelnen Substantive mit 62,30% aus, haben wir es doch sowohl bei den Einträgen in den fünf Registern wie auch bei den extrahierten Nominalgruppen in den meisten Fällen mit Substantiven zu tun. In diese Kategorie fallen auch viele Substantive, die keine Schlagwörter sind, sondern in geläufigen Formulierungen oder in den Fachgebieten besonders häufig verwendet werden. Beispiele sind: „Abschnitt“, „Bezeichnung“, „Form“, „Satz“, „Sprache“, „Unterschied“ usw. 16,98% des Ballasts betreffen die Kombination Substantiv plus attributives Adjektiv. Dieser Anteil ist etwas kleiner als derjenige bei den Treffern (20,98%). Der Anteil der Schlagwörter, die aus zwei Substantiven gebildet werden, ist mit 12,11% erstaunlich hoch, in Anbetracht der Komplexität vieler Nominal- oder Präpositionalphrasen und der Vielfältigkeit der Ausdrucksweisen in den Textausschnitten

jedoch einleuchtend. In allen gefundenen nicht relevanten Wortgruppen ist das Ordnungswort ein Substantiv; ein Umstand, der auf die Extraktionsmethode und die hergestellte kanonische Wortreihenfolge der Wortgruppen zurückzuführen ist. 63,17% beträgt der Anteil der einzelnen Wörter, 36,83% der Multiwortterme bezüglich des Ballastes. Auch dieses Verhältnis hat mit der Verwendung vieler Substantive in geläufigen Formulierungen zu tun sowie mit der Zerlegung der extrahierten Phrasen in alle ihre untergeordneten Nominal- und Präpositionalphrasen. Man müsste vielleicht grundsätzlich überprüfen, ob sich mit der ausschliesslichen Verwendung der übergeordneten Phrasen die besseren oder schlechteren Resultate ergeben würden. Die Entscheidung zur vollständigen Zerlegung entstand aus der Situation heraus, dass viele aus den Textausschnitten extrahierten Nominal- und Präpositonalphrasen von beträchtlicher Länge und Komplexität waren mit bis zu fünf Subordinierungen.

Auch das Ergebnis des erzeugten Ballastes ist enttäuschend. 73,69% aller gefundenen nominalen Wortgruppen wurden fälschlicherweise als Indexterme zugewiesen (b dividiert durch a+b).

5.3.3 Zusammenfassung

Durch die Untersuchung sowie Indexierung der Textkörper und den Vergleich mit den dazugehörigen Buchregistern können wir feststellen, dass das Extraktionsverfahren mit der Gewichtung der Terme durch die inverse Seitenhäufigkeit nicht zufrieden stellende Ergebnisse liefert, welche eine direkte und isolierte Umsetzung des Verfahrens für eine automatische Indexierung von deutschsprachigen Texten unausführbar machen. Ein Recall von 0,26 sowie eine Precision von 0,26 bezeugen ein schlechtes Gesamtergebnis und lassen eine Bewertung des Indexierungsverfahrens ernüchternd ausfallen.

Insbesondere bereitet die Ermittlung und Zuteilung der Mehrwortterme Schwierigkeiten. Zusätzlich können Einträge, die als Ordnungswort eine andere lexikalische Wortkategorie als ein Substantiv besitzen, mit der Methode unmöglich identifiziert werden. Im Kapitel 7 „Ausblick“ werde ich auf diese Probleme und ihre eventuelle Beseitigung nochmals zu sprechen kommen.

Damit ist der statistische Teil der Arbeit beendet.

6 Zusammenfassung und Schlussfolgerungen

Die vorliegende Lizentiatsarbeit „Das Indexieren von natürlichsprachlichen Dokumenten und die inverse Seitenhäufigkeit“ beschäftigte sich mit dem Indexieren von Dokumenten, Texten oder Büchern, welche in natürlicher Sprache verfasst wurden. Das Ziel war, eine möglichst umfassende Übersicht über die verschiedenartigen Merkmale und Funktionen von Indexen sowie die zahlreichen Verfahren und Techniken des Indexierens und auch des Evaluierens von Indexierungsergebnissen zu geben. Zusätzlich sollte eine selbst entwickelte Methode für die Indexierung von deutschsprachigen Texten auf ihren Gebrauchswert hin anhand von existierenden Büchern mit Schlagwortverzeichnissen getestet werden.

Im ersten, theoretischen Teil der Arbeit wurde das Fachgebiet des Indexierens vorgestellt. Es ging vorerst um die verschiedenen Indextypen und Ebenen des Indexierens, um die Definition und die Funktion eines Index. Wir haben erfahren, dass eine Indexierung schriftlich oder bildhaft niedergelegte Informationen erschliesst, um diese in einem späteren Bedarfsfall wieder auffindbar zu machen. Beim Prozess des Indexierens werden einem Originaltext inhaltskennzeichnende Begriffsbenennungen, sogenannte Indexterme, zugewiesen. Der daraus resultierende Index ist eine Textrepräsentation, welcher den thematischen Inhalt der Originaldokumente mit einem bestimmten Grad an Detailliertheit verdichtet. Wir haben Autorenindexe, Schlagwortregister, klassifizierte und koordinierte Indexe, permutierte Titelindexe, facettierte Indexe, Kettenindexe, Zeichenkettenindexe sowie Zitierindexe differenziert. Ein wesentlicher Faktor beim Indexieren spielt die Indexsprache. Das Vokabular einer Indexsprache kann dem Originaldokument entnommen, frei formuliert oder kontrolliert sein. Die häufigsten Hilfsmittel beim Indexieren mit einem kontrollieren Wortschatz sind Thesauri und Klassifikationen, bisweilen auch Schlagwortlisten. Natürlichsprachliche oder freie Indexterme haben den Vorteil, dass sie ausdrucksstark und flexibel sind. Die Indexterme einer kontrollierten Indexsprache dagegen lösen Mehrdeutigkeiten auf, lassen die Suche nach Allgemeinbegriffen zu und verbinden zusammenhängende Indexterme miteinander. Allerdings ist der Aufwand für die Erstellung und die Pflege eines kontrollierten Vokabulars nicht zu unterschätzen. Durch die Verwendung einer Indexsprachengrammatik können Relationen zwischen den Indextermen ausgedrückt werden. Eine solche Indexierung nennt man strukturiert oder syntaktisch. Bei für den Druck bestimmten Schlagwortregistern können Indexsprachengrammatiken nicht verwendet werden. Für sie ist eine Strukturierung mittels Präkombination wesentlich.

Im Abschnitt 2.5 wurden die zahlreichen Verfahren, Methoden und Techniken des Indexierens aufgezeigt. Eine Indexierung kann manuell-intellektuell oder automatisch-maschinell durchgeführt werden. Daneben existiert auch die computerunterstützte Indexierung, bei welcher ein automatisches Verfahren Indexterme vorschlägt, die in einer manuellen Nachbearbeitung bestätigt werden. Sowohl manuelle wie auch automatische Indexierung gehen mit Extraktionsmethoden oder Zuteilungsmethoden vor. Beim intellektuellen Indexieren ist die Extraktionsmethode eher selten, bei den manuellen Zuteilungsmethoden unterschieden wir freies, kontrolliertes, klassifizierendes

sowie verbindliches Indexieren. Das manuelle hybride Indexieren kombiniert freies und kontrolliertes Indexieren. Kontrastierend dazu wenden die meisten automatischen Indexierungsverfahren Extraktionsmethoden an. Mittels linguistischen Verfahren werden Dokumente vorverarbeitet (Stemmatisierung, Lemmatisierung, Normalisierung, Phrasenerkennung, Stoppwortentfernung), um danach mit statistischen oder kollektionsorientierten Methoden weiterverarbeitet zu werden. Die in der Praxis eingesetzten Verfahren basieren vor allem auf der Analyse der Worthäufigkeit. Für die in den Dokumenten vorkommenden natürlichsprachlichen Wörter werden Gewichte berechnet, die auf der Annahme basieren, dass ein häufig auftretendes Wort in einem Dokument für das Thema des Dokuments von Wichtigkeit ist. Zur Berechnung der Gewichte gibt es zahlreiche Formeln. Die am häufigsten verwendete ist diejenige der inversen Dokumenthäufigkeit. Sie bildete auch die Basis für die Entwicklung der inversen Seitenhäufigkeit im zweiten Teil der Arbeit. Die starke statistische bzw. probabilistische Orientierung führt bedauerlicherweise oft dazu, dass inhaltsrelevante Terme nicht immer im gewünschten Masse berücksichtigt werden. Als Abschluss des Kapitels zu den Indexierungsverfahren haben wir die verschiedenen Verfahren einem Vergleich unterzogen und festgestellt, dass dem Kriterium Wirtschaftlichkeit vermehrt eine grosse Bedeutung beigemessen wird. Die gegenwärtige Situation, in der wir mit riesigen Mengen von elektronischen und anderen natürlichsprachlichen Dokumenten und Texten konfrontiert werden, bedarf einer automatischen Indexierung. Das manuelle oder intellektuelle Indexieren wird immer weniger tragbar, es ist zu zeit- und kostenaufwändig. Kaiser ist sogar der Ansicht, dass die manuelle Indexierung an der immer grösser werdenden Informationsflut hinsichtlich der zeitlichen und fachlichen Belastung des Indexierers scheitern wird.[293] Neue Medien wie Bild und Ton, die zunehmend in Dokumente integriert werden, erschweren die manuelle Indexierung zusätzlich. Auch Moens bestätigt das nicht mehr länger realisierbare manuelle Indexieren und die grosse Nachfrage nach Systemen zur automatischen Indexierung. Aber sie beurteilt die gegenwärtigen Systeme oder Versuche als nicht brauchbar, insbesondere bei der Indexierung von Büchern.[294] Alternative Techniken müssen also gefunden werden und haben sich auch entwickelt (z.B. Volltextspeicherung, Relevance-Feedback). Es scheint mir dabei wesentlich, dass ein Entscheid für eines der sehr vielen unterschiedlichen Indexierungsverfahren mit ganz verschiedenen Ansätzen immer den Zweck und das Ziel der Indexierung vor Augen haben sollte und dass ein aussagekräftiger Vergleich der Verfahren zweckorientiert erfolgen muss.

Den Schlagwortregistern widmete ich ein eigenes Kapitel, sind sie doch das Ziel des Indexierungsverfahrens im statistischen Teil. Die Bestandteile eines Buchindex bilden Schlagwörter, Untereinträge, Fundstellen und Querverweise. Wir unterschieden Namen- und Schlagwortregister sowie alphabetische und systematische Verzeichnisse. Um eventuell Erkenntnisse für das automatische Erstellen von Schlagwortregistern gewinnen zu können, wurde das manuelle Vorgehen beim Indexieren eines Buches demonstriert. Ich möchte diesbezüglich festhalten, dass für eine automatische Indexierung die intellektuellen Prozesse nicht einfach imitiert werden können und dass andere Methoden benötigt werden. Diese These wird auch von Kuhlen

[293] Kaiser 1996:48
[294] Moens 2000:227

unterstützt.[295] Schliesslich betrachteten wir auch Aspekte der Form und des Layouts von Registern. Im Mittelpunkt standen dabei stets die Nützlichkeit eines Registers sowie die Korrektheit und Konsistenz.

Im Kapitel 4 „Evaluation eines Index“ wurde aufgezeigt, wie Vergleiche und Bewertungen von Indexierungen vorgenommen werden. Es wurden Merkmale von „guten“ Indexen aufgeführt und Faktoren aufgelistet, welche eine Indexierung beeinflussen können. Zur direkten Bewertung von Indexen werden üblicherweise die vier Bewertungskriterien Exhaustivität, Spezifität, Korrektheit und Konsistenz hinzugezogen. Werden Indexierungsresultate eines einzigen Indexierers verglichen, so wird die Intra-Indexiererkonsistenz geprüft. Die Inter-Indexiererkonsistenz wägt die Konsistenz verschiedener Indexierer gegeneinander ab. Die indirekten Bewertungen erfolgen mittels Retrievaltests, die der Berechnung der Performanz eines Informationssystems dienen. Sie kamen im zweiten Teil der Arbeit zur Anwendung und basieren auf der Annahme: Je effektiver ein System, desto besser befriedigt es die Benutzerbedürfnisse. Wir haben die Methoden Fehlerstatistik und den Konsistenzfaktor q kennen gelernt. Die weitaus am häufigsten verwendeten Masse zur Bewertung der Performanz eines Informationssystems sind Recall und Precision. Sie stützen sich auf das grundlegende Bewertungskriterium der Relevanz ab. Relevanz ist keine unumstrittene Grösse; man unterscheidet deshalb Benutzer- und Systemrelevanz. Weitere Masse zur Messung der Retrievaleffektivität sind Fallout, Accuracy, Error, E-Wert und F-Wert. Ausser den zwei letztgenannten wurden alle Bewertungsmasse bei der Prüfung der mit der inversen Seitenhäufigkeit erzielten Resultate hinzugezogen. Für die meisten Systeme zur automatischen Indexierung gilt es, möglichst hohe Recall- und Precisionwerte zu erreichen.

Der zweite Teil der Arbeit bestand in statistischen Untersuchungen. Die Register von fünf realen Büchern wurden hinsichtlich der Grösse, Anzahl Tokens, der lexikalischen Wortkategorien, der Anzahl Schlagwörter und Untereinträge sowie der lexikalischen und syntaktischen Bestandteile der Einträge erforscht. Es hat sich herausgestellt, dass die Kategorie der Substantive in allen analysierten Registern am häufigsten vertreten ist und einen Anteil von 65,55% aller Wortformen ausmacht. Wir konnten auch erfahren, dass alle Register zusammen im Schnitt 2,54 Einträge pro Buchseite zur Verfügung stellen, was wesentlich weniger ist, jedoch intuitiv angemessener als die von Moens geforderten 6 bis 8 Einträge pro Seite. Jedes Schlagwort der fünf Register besteht durchschnittlich aus 1,6 Tokens und jeder Untereintrag sogar aus 1,83 Tokens. Dies bedeutet, dass die Mehrwortterme in den Schlagwortverzeichnissen eine wesentliche Rolle spielen und dass mit mehrgliedrigen Indexeinträgen präzise Beschreibungen erreicht werden können. In Prozenten ausgedrückt bestehen 38,81% aller Einträge aus einem Mehrwortterm und entsprechend 61,19% aus einem einzelnen Wort, wobei 94,30% aller Einzelworteinträge Substantive sind. Insgesamt setzten sich die Mehrheit der Einträge der fünf Register aus einem Substantiv als Schlagwort und einem Adjektiv als nähere Bestimmung zusammen.

[295] Kuhlen 1986:135

Ein jeweils 50-seitiger Ausschnitt aus den zu den fünf Registern dazugehörigen Textkörpern wurde ebenfalls einer Untersuchung unterzogen und automatisch indexiert, um den Gebrauchwert der Indexierungsmethode der inversen Seitenhäufigkeit zu testen. Die inverse Seitenhäufigkeit ist eine von mir entwickelte Methode zur Gewichtung von Wörtern oder Wortgruppen im Hinblick auf das Indexieren von Büchern. Sie basiert auf der inversen Dokumenthäufigkeit, welche für die Indexierung von Dokumenten in grossen Dokumentkollektionen verwendet wird. Aus den ausgewählten Textausschnitten wurden zunächst unter Berücksichtigung der Tatsache, dass in den Schlagwortverzeichnissen zahlreiche Multiwortterme vorkommen, die Nominal- und Präpositonalphrasen extrahiert und aus diesen auch alle untergeordneten Nominal- und Präpositionalphrasen. Sämtliche Funktionswörter wurden eliminiert, die Phrasen lemmatisiert und auf eine einheitliche Form gebracht. Für jede verbleibende Wortgruppe konnte nun die inverse Seitenfrequenz berechnet werden. Diejenigen Wortgruppen, welche ein Gewicht von grösser/gleich 1,71 erzielten, ergaben die gefundenen Indexterme. Die in den Registern stehenden Indexterme bildeten die relevanten Indexterme. Auf diese Weise konnten die Indexterme in gefundene relevante, gefundene nicht relevante, relevante nicht gefundene und nicht relevante nicht gefundene Indexterme eingeteilt und die der Untersuchungssituation angepassten Bewertungsmasse Recall, Precision, Fallout, Accuracy sowie Error berechnet werden.

Der in allen fünf Büchern erzielte durchschnittliche Recallwert betrug 0,26, ebenso der Wert der Precision. Der Fallout oder Ballast belief sich auf 0,06. Es konnte eine Treffgenauigkeit von 0,90 und eine Fehlerrate von 0,10 bei einem minimalen Grenzwert von 1,71 erlangt werden. Während die Werte der Ausbeute und der Präzision auf eine ungenügende und enttäuschende Performanz der Gewichtungsmethode der inversen Seitenhäufigkeit hinweisen, so ist der Ballastanteil akzeptierbar und die Genauigkeit bei den Treffern einigermassen zufrieden stellend. Eine Heraufsetzung des minimalen Grenzwertes auf 2,01 brachte keine besseren Ergebnisse, im Gegenteil, sie verschlechterten sich in Bezug auf Recall, Precision und Fallout.

Die statistischen Untersuchungen haben demonstriert, dass sich mit der inversen Seitenhäufigkeit keine zufrieden stellende automatische Indexierung von deutschsprachigen Texten oder Büchern durchführen lässt. Insbesondere konnte gezeigt werden, dass der ausschliessliche Einsatz der einfachen linguistischen Methoden sowie der Termgewichtungsmethode der inversen Seitenfrequenz zu keinen befriedigenden Ergebnissen führt und verbesserungsfähig ist. Die ursprünglich formulierte Hypothese, dass mit der Methode ein Zusammenhang zwischen einem Buchtext und dem dazugehörigen Schlagwortregister hergestellt werden könne, ist somit nur ansatzweise erfüllt worden. Dennoch bin ich der Überzeugung, dass eine verfeinerte Variante des Verfahrens in Kombination mit anderen Methoden bessere Ergebnisse liefern kann. Auf diese zusätzlichen Methoden werde ich nun im folgenden letzten Abschnitt meiner Lizentiatsarbeit noch kurz eingehen.

7 Ausblick

Eine zukünftige Verwendung der inversen Seitenhäufigkeit sollte meines Erachtens verstärkt die Einbindung von zusätzlichen linguistischen Methoden verfolgen. Hierbei könnten folgende Ansätze in Betracht gezogen werden:

- Nutzung von lexikalischen und syntaktischen Informationen: Bestimmte Wörter oder Wortgruppen wie z.B. „die Bezeichnung“, „der Begriff“, „Definition“, „sogenannte“, „nennt man“, „bezeichnet man“ usw. könnten Hinweise für Schlagwortkandidaten geben. Dazu müssten die Schlagwörter der bestehenden fünf Register in den Texten aufgesucht werden, um zu ermitteln, in welchen lexikalischen und/oder syntaktischen Kontext sie eingebettet sind. Aus diesen Kontexten lassen sich möglicherweise verschiedene linguistische Muster entwickeln, die Indizien für Schlagwortkandidaten bilden, um vor allem den Recall und auch die Precision zu verbessern.

- Nutzung von Diskursstrukturen: Den Wortformen der Titel, Topicsätze sowie Hauptsätze sollte grössere Bedeutung bei der Termgewichtung zukommen. Sie stellen Themen und Gegenstände in den Vordergrund, während in Texten üblicherweise Nebensätze zusätzliche und periphere Dinge beschreiben. Auch Abschnitte mit Überschriften wie „Begriffsdefinitionen“, „Begriffsvereinbarungen“ oder Abstracts könnten auf diese Art eine Bevorzugung erlangen.

- Verwendung von Phrasenköpfen: Da die aus den Textausschnitten extrahierten nominalen Wortgruppen teilweise sehr komplex aufgebaut und reich in ihren Formulierungsausgestaltungen sind, werden nur die Phrasenköpfe als Schlagwortkandidaten verwendet. Die Idee muss allerdings zuerst überprüft werden.

- Bedingte Zerlegung der übergeordneten Phrasen: Nur bei Nominal- und Präpositionalphrasen, die mehr als beispielsweise zwei Einbettungsstufen aufweisen, soll eine Extraktion der untergeordneten Phrasen durchgeführt werden. Möglicherweise lässt sich so die Anzahl der gefundenen nicht relevanten Indexterme verkleinern; auch diese Methode ist auf jeden Fall vorab zu testen.

- Kompositazerlegung: Zur Erreichung eines höheren Recalls könnte auch eine Verwendung der einzelnen Bestandteile von Komposita als Indexterme beitragen. Ein Vergleich von extrahierten Komposita mit Indexeinträgen muss dafür zunächst gezogen werden.

- Verwendung von Funktionswörtern: Bei mehrgliedrigen Wortgruppen werden die Funktionswörter, im Speziellen die Präpositionen, wieder einfügt, um die Bezüge und Verhältnisse innerhalb der Wörter eines Schlagwortkandidaten zu klären. Auch damit sollte sich die Menge des Ballasts reduzieren lassen.

Nicht nur linguistische, sondern auch auf formalen Aspekten beruhende Methoden betrachte ich als Möglichkeit, das Termgewichtungsverfahren effizienter zu gestalten:

- Nutzung typografischer Informationen: Fettdruck, Kursivdruck oder Anführungs- und Schlusszeichen werden als Signale für mögliche Indexterme gewertet, um die Ausbeute sowie die Präzision der Indexierung zu erhöhen.

- Mehrere Seiten als Fundstelle: Tritt eine nominale Wortgruppe in einem Buchtext über mehrere nacheinanderfolgende Seiten hinweg auf, das Gewicht fällt aber für jede einzelne Buchseite nicht hoch genug aus, um die Wortgruppe als Indexterm zu identifizieren, so könnte man versuchen, eine einzige Termgewichtung der Wortgruppe für eine Folge von Buchseiten anzustellen. Fundstellen wie „35-38“ könnten erfasst werden, der Recall könnte sich möglicherweise verbessern.

Bisher wurden die Möglichkeiten der automatischen Textkategorisierung nicht in die statistische Indexierung mit einbezogen (vergleiche Abschnitt 2.5.2.2.2). Durch eine ausgiebige Erforschung des Gebietes lassen sich vielleicht wichtige Anhaltspunkte zur Verbesserung einer automatischen Indexierung finden. Ebenso sind auch Verfahren und Techniken aus anderen Informationsverarbeitungsgebieten für das Indexieren von Texten denkbar:

- Methoden der Termextraktion: Mithilfe der automatischen Termextraktion wird die Fachterminologie eines Sachgebietes in Dokumenten erkannt. Ganz grundsätzlich könnte eine erfolgreiche Termextraktion für das maschinelle Indexieren von grossem Nutzen sein. Bedauerlicherweise sind die Erfolge auf diesem Gebiet gegenwärtig eher bescheiden.[296] Eine der ursprünglichen Ideen der vorliegenden Arbeit war, neben dem Verfahren der Termgewichtung auch eine automatische Termextraktion mit einer bestehenden Software durchzuführen. Aufgrund der Erkenntnisse in den Arbeiten von Heidemann/Volk und Niederbäumer wurde die Idee jedoch aufgegeben.

- Methoden des Textmining: Die Techniken des Textmining und der Informationsextraktion liessen sich zum Beispiel für die Eigennamenerkennung oder die Identifizierung von linguistischen Indizien für Schlagwortkandidaten und somit für eine verbesserte automatische Indexierung verwenden.

- Neue Formen der Wissensrepräsentation: Reimer hält in seinen Ausführungen „Neue Formen der Wissensrepräsentation“ fest, dass die traditionellen Indexierungsverfahren (zu welchen die hier angewendeten auch gehören) nicht in der Lage sind, über eine rein syntaktische Gruppierung von

[296] Heidemann/Volk 1999; Niederbäumer 2000

Wortgruppen hinaus die textspezifischen Beziehungen zwischen den Indextermen zu erfassen und in einen Index aufzunehmen. Er schlägt vor, auf andere terminologische Repräsentationsformalismen für thematische Dokumentbeschreibungen auszuweichen: semantische Netze oder terminologische Logiken.[297] Allerdings gehen diese neuen Formalismen von einem komplett andersartigen Ansatz aus und sind nicht auf einfache Weise in das bisherige Verfahren der Termgewichtung zu integrieren. Die Idee müsste detaillierter verfolgt werden.

Vielleicht besteht zusätzlich die Möglichkeit, das zur Zeit in der Abteilung Computerlinguistik der Universität Zürich erstellte „Glossar der Computerlinguistik“ als kontrolliertes Vokabular neben dem extrahierten für die Indexierung der Vorlesungsskripte zu nutzen; ein fachspezifischer Thesaurus steht momentan nicht zur Verfügung.

Die Idee der Gewichtungsmethode der inversen Seitenhäufigkeit hoffe ich unter Berücksichtigung möglichst vieler der eben erwähnten Punkte (vor allem bezüglich der linguistischen Methoden) weiter verfolgen zu können. Ein wesentlicher Schritt der zukünftigen Arbeiten wird die vollständige Implementierung des gesamten Verfahrens sein. Für die vorliegenden statistischen Untersuchungen sowie die automatische Indexierung wurden jeweils nur die einzelnen Verarbeitungsschritte maschinell durchgeführt und nachträglich manuell kontrolliert. Eine Programmierung des gesamten Verfahrens bildet die Voraussetzung, um auch grössere Textausschnitte oder ganze Buchtexte erfassen und mit den dazugehörigen Schlagwortregistern vergleichen zu können. Nicht zuletzt wird davon abhängen, ob die grundsätzliche Idee der inversen Seitenhäufigkeit zu befriedigenden Indexierungsergebnissen führen kann.

[297] Reimer 1997:180-198

8 Literaturverzeichnis

8.1 Untersuchte Bücher

[Cap 1993] Cap, C. H.: Theoretische Grundlagen der Informatik. Wien/New York: Springer-Verlag 1993.

[Hausser 2000] Hausser, R.: Grundlagen der Computerlinguistik. Mensch-Maschine-Kommunikation in natürlicher Sprache. Berlin/Heidelberg etc.: Springer-Verlag 2000.

[Linke et al. 1996] Linke, A./Nussbaumer, M./Portmann, P. R.: Studienbuch Linguistik. Ergänzt um ein Kapitel „Phonetik und Phonologie“ von Urs Willi. 3., unveränderte Auflage (1991). Tübingen: Max Niemeyer Verlag 1996. (= Reihe Germanistische Linguistik; Kollegbuch 121)

[Stegmüller 1986] Stegmüller, W.: Hauptströmungen der Gegenwartsphilosophie. Eine kritische Einführung. Band 2. 7., erweiterte Auflage (1952). Stuttgart: Alfred Kröner Verlag 1986. (= Kröners Taschenausgabe; Band 309)

[Werlen 1989] Werlen, I.: Sprache, Mensch und Welt. Geschichte und Bedeutung des Prinzips der sprachlichen Relativität. Darmstadt: Wissenschaftliche Buchgesellschaft 1989. (= Erträge der Forschung; Band 269)

8.2 Sekundärliteratur

[Aas/Eikvil 1999] Aas, K./Eikvil, L.: Text Categorisation: A Survey. Norwegian Computing Center, Oslo. June 1999. http://www.nr.no/samba/textmining.html

[American Society of Indexers 2000] American Society of Indexers: Bibliography. Wheat Ridge, CO. Oktober 2000. http://www.asindexing.org/bibliog.shtml

[Belew 2000] Belew, R. K.: Inverse Document Frequency. University of California, Computer Science and Engineering Department, San Diego, CA. 21.09.2000. http://www.cse.ucsd.edu/~rik/foa/l2h/foa-3-3-7.html

[Bhattacharyya 1982] Bhattacharyya, G.: Classaurus: Its Fundamentals, Design, and Use. Unterlagen zur Tagung „INDEKS 1992“ der Gesellschaft für Klassifikation. Frankfurt 1982. Absatz 321. (= Studien zur Klassifikation; Band 11)

[Bonura 1994] Bonura, L. S.: The Art of Indexing. New York etc. 1994. (= Wiley Technical Communication Library)

[Borko/Bernier 1978] Borko, H./Bernier, C. L.: Indexing Concepts and Methods. New York/ London 1978.

[Burkart 1997] Burkart, M.: Thesaurus. In: Buder, M. et al. (Hgg.): Grundlagen der praktischen Information und Dokumentation. Ein Handbuch zur Einführung in die fachliche Informationsarbeit. 4. völlig neu gefasste Ausgabe (1972). München etc. 1997. S. 160-179.

[Chang et al. 1992] Chang, J.-S. et al.: A Corpus-based Statistical Approach to Automatic Book Indexing. Proceedings of “The Third Conference on Applied Natural Language Processing”. Trento (Italien) 1992. S. 147-151.

[Chung et al. 1998] Chung, Y.-M./Pottenger, W. M./Schatz, B. R.: Automatic Subject Indexing Using an Associative Neural Network. University of Illinois, CANIS – Community Systems Lab, Champaign, IL. 05.09.1998. http://www.canis.uiuc.edu/projects/interspace/technical/canis-report-0003/

[Cleveland/Cleveland 1990] Cleveland, D. B./Cleveland, A. D.: Introduction to Indexing and Abstracting. Second Edition (1983). Englewood, CO 1990.

[Dillon/Gray 1983] Dillon, M./Gray, A. S.: FASIT: A Fully Automatic Syntactically Based Indexing System. In: Journal of the American Society for Information Science, 34 (2), 1983. S. 99-108.

[Fagan 1988] Fagan, J. L.: Experiments in Automatic Phrase Indexing for Document Retrieval. Ann Arbor, MT 1988.

[Fagan 1989] Fagan, J. L.: The Effectiveness of a Nonsyntactic Approach to Automatic Phrase Indexing for Document Retrieval. In: Journal of the American Society for Information Science, 40 (2), 1989. S. 115-132.

[Ferber 2000] Ferber, R.: Data Mining und Information Retrieval. Online Kurs an der FernUniversität Hagen. FernUniversität, Fachbereich Elektrotechnik und Informationstechnik, Hagen. 14. 01. 2000. http://teefix.fernuni-hagen.de/~ferber/kurse/dm-ir/vl/book_1.part_4.chapter_3.section_1.subdiv1_3.html#192

[Frei 1999] Frei, H. P.: BewSuchtes. Unterlagen zur Vorlesung „Informationssuche in heterogenen Datensammlungen im Web“. Zürich, Universität Zürich, Abteilung Wirtschaftsinformatik. Vorlesung vom 11.09.1999.

[Fugmann 1999] Fugmann, R.: Inhaltserschliessung durch Indexieren: Prinzipen und Praxis. Frankfurt am Main 1999. (= Reihe Informationswissenschaft der DGD [Deutsche Gesellschaft für Dokumentation e.V.]; Band 3)

[Gallmann/Sitta 1996] Gallmann, P./Sitta, H.: Deutsche Grammatik. Zürich 1996.

[Godby 1994] Godby, J.: Two Techniques for the Identification of Phrases in Full Text. Dublin, OH 1994. http://www.oclc.org/oclc/research/publications/arr/1994/part1/twotech.htm

[Goeser 1997] Goeser, S.: Inhaltsbasiertes Information Retrieval: Die TextMining-Technologie. Fachhochschule Darmstadt, Fachbereich IuD, Darmstadt. 16.03.1998 http://www.iud.fh-darmstadt.de/iud7wwwmeth/publ/paper/ldvf97/goeser1.htm

[Graf 1995] Graf, P.: Term Indexing. Berlin 1995.

[Hahn 1986] Hahn, U.: Methoden der Volltextverarbeitung in Informationssystemen. Ein State-of-the-Art-Bericht. In: Kuhlen, R.: Informationslinguistik. Theoretische, experimentelle, curriculare und prognostische Aspekte einer informationswissenschaftlichen Teildisziplin. Tübingen 1986. S. 195-216. (= Sprache und Information; Band 15)

[Harter 1986] Harter, S. P.: Online Information Retrieval: Concepts, Principles, and Techniques. San Diego, CA 1986.

[Hauck 1997] Hauck, C.: Quantität und Qualität von Fachbuchregistern auf dem Gebiet der elektronischen Datenverarbeitung. Abschlussarbeit im Weiterbildungsstudium „Wirtschafts- und Fachinformation“. Technische Universität Ilmenau, Ilmenau 1997.

[Heidemann/Volk 1999] Heidemann, B./Volk, M.: Evaluation of Terminology Extraction Tools. Zürich 1999.

[Hutchins 1985] Hutchins, W. J.: Information Retrieval and Text Analysis. In: van Dijk, T.A. (Hg.): Discourse and Communiation: New Approaches to the Analysis of Mass Media Discourse and Communication. Berlin 1985. S. 106-125.

[Jonák 1984] Jonák, Z.: Automatic Indexing of Full Texts. In: Information Processing & Management, 20 (5/6), 1984. S. 619-627.

[Kaiser 1996] Kaiser, A.: Computer-unterstütztes Indexieren in intelligenten Information Retrieval Systemen. Ein Relevanz-Feedback orientierter Ansatz zur Informationserschliessung in unformatierten Datenbanken. Dissertation. Wirtschaftsuniversität Wien, Abteilung Informationswissenschaft, Wien. 20.08.1998. http://wwwai.wu-wien.ac.at/Publikationen/Kaiser/diss.html

[Klenke 1999] Klenke, M.: Methoden der automatisierten Informationsextraktion durch rechnergestützte Klassifikation von Fernerkundungsdaten. Dissertation. Friedrich-Schiller-Universität Jena, Chemisch-Geowissenschaftliche Fakultät, Jena. 04.11.1999. http://home.germany.net/100-311822/ www_diss/ mk_node19.html

[Knight 1979] Knight, G. N.: The Art of Indexing: A Guide to the Indexing of Books and Periodicals. London 1979.

[Knorz 1983] Knorz, G.: Automatisches Indexieren als Erkennen abstrakter Objekte. Tübingen 1983. (= Sprache und Information; Band 8)

[Knorz 1997] Knorz, G.: Indexieren, Klassieren, Extrahieren. In: Buder, M. et al. (Hgg.): Grundlagen der praktischen Information und Dokumentation. Ein Handbuch zur Einführung in die fachliche Informationsarbeit. 4. völlig neu gefasste Ausgabe (1972). München etc. 1997. S. 120-140.

[Kuhlen 1986] Kuhlen, R.: Some Similarities and Differencies between Intellectual and Machine Text Understanding for the Purpose of Abstracting. In: Kuhlen, R.: Informationslinguistik. Theoretische, experimentelle, curriculare und prognostische Aspekte einer informationswissenschaftlichen Teildisziplin. Tübingen 1986. (= Sprache und Information; Band 15)

[Lahtinen 2000] Lahtinen, Timo: Automatic Indexing: An Approach Using an Index Term Corpus and Combining Linguistic and Statistical Methods. Dissertation. University of Helsinki, Faculty of Arts, Department of General Linguistics, Helsinki 2000. http://ethesis.elsinki.fi/ulkaisut/hum/yleis/vk/lahtinen/automati.pdf

[Lancaster 1986] Lancaster, F. W.: Vocabulary Control for Information Retrieval. Second edition (1976). Arlington, VA 1986.

[Lancaster 1998] Lancaster, F. W.: Indexing and Abstracting in Theory and Practice. Second edition (1991). London 1998.

[Larson/Hearst 1998] Larson, R./Hearst, M.: Term Weighting and Ranking Algorithms. University of California, School of Information Management and Systems, Berkeley, CA. 15.10.1998. http://www.sims.berkeley.edu/courses/is202/f98/Lecture17/sld001.htm

[Lustig 1986] Lustig, G.: Automatische Indexierung zwischen Forschung und Anwendung. Hildesheim 1986.

[Manecke 1997] Manecke, H.-J.: Klassifikation. In: Buder, M. et al. (Hgg.): Grundlagen der praktischen Information und Dokumentation. Ein Handbuch zur Einführung in die fachliche Informationsarbeit. 4. völlig neu gefasste Ausgabe (1972). München etc. 1997. S. 141-159.

[Miller 1997] Miller, U.: Thesaurus Contruction: Problems and their Roots. In: Information Processing & Management, 33 (4), 1997. S. 481-493.

[Moens 2000] Moens, M.-F.: Automatic Indexing and Abstracting of Document Texts. Boston/Dordrecht/London 2000. (= The Kluwer International Series on Information Retrieval; 6)

[Mulvany 1994] Mulvany, N. C.: Indexing Books. Chicago/London 1994. (= Chicago Guides to Writing, Editing, and Publishing)

[Neumann 2000] Neumann, G.: Informationsextraktion. Deutsches Forschungszentrum für Künstliche Intelligenz, Saarbrücken. 21.11.2000. http//www.dfki.de/~neumann/publications/new-ps/ie.pdf

[Niederbäumer 2000] Niederbäumer, A.: German Terminology of Banking: Linguistic Methods of Description and Implementation of a Program for Term Extraction. Lizentiatsarbeit. Universität Zürich, Abteilung Computerlinguistik, Zürich 2000.

[Reimer 1997] Reimer, U.: Neue Formen der Wissensrepräsentation. In: Buder, M. et al. (Hgg.): Grundlagen der praktischen Information und Dokumentation. Ein Handbuch zur Einführung in die fachliche Informationsarbeit. 4. völlig neu gefasste Ausgabe (1972). München etc. 1997. S. 180-207.

[van Rijsbergen 1979] van Rijsbergen, C. J.: Information Retrieval. Second edition (1975). London 1979. http://www.ifi.unizh.ch/CL/hess/classes/ecl1/rijsbergen.pdf

[Rowley 1988] Rowley, J. E.: Abstracting and Indexing. Second edition (1982). London 1988.

[Ruge 1995] Ruge, G.: Wortbedeutung und Termassoziation: Methoden zur automatischen semantischen Klassifikation. Hildesheim/Zürich/New York 1995. (= Sprache und Computer; Band 14)

[Salton 1989] Salton, G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Reading, MA etc. 1989.

[Schneider 2001] Schneider, G.: Die Anwendung von Morphologieanalyse in Information Retrieval-Systemen. Vorlesungsunterlagen. Universität Zürich, Abteilung Computerlinguistik, Zürich. 12.06.2001. http://www.ifi.unizh.ch/CL/gschneid/LexMorphVorl/Lexikon10.IR.pdf

[Schwantner 1991] Schwantner, M.: Aufbau und Pflege eines Wörterbuches für die automatische Indexierung. Darmstadt 1991.

[Seeger 1997] Seeger, T.: Grundbegriffe der Information und Dokumentation. In: Buder, M. et al. (Hgg.): Grundlagen der praktischen Information und Dokumentation. Ein Handbuch zur Einführung in die fachliche Informationsarbeit. 4. völlig neu gefasste Ausgabe (1972). München etc. 1997. S. 1-15.

[Smeaton 1997a] Smeaton, A.: Information Retrieval: Still Butting Heads with Natural Language Processing. Dublin City University, School of Computer Applications, Dublin 1997. http://www.ifi.unizh.ch/CL/hess/classes/ecl1/smeaton.CA2197.pdf

[Smeaton 1997b] Smeaton, A.: Using NLP or NLP Resources for Information Retrieval Tasks. Dublin City University, School of Computer Applications, Dublin 1997. http://www.ifi.unizh.ch/CL/hess/classes/ecl1/smeaton.CA2297.pdf

[Sparck-Jones/Galliers 1995] Sparck-Jones, K./Galliers, J.R.: Evaluating Natural Language Processing Systems. An Analysis and Review. Berlin etc. 1996. (= Lecture Notes in Artificial Intelligence; 1083)

[Steiger 1982] Steiger, R.: Die syntaktisch-semantische Textanalyse und automatisierte Indexierung. Leipzig 1982. (= Einführung in die Information und Dokumentation; 14)

[Verdejo et al. 1999] Verdejo, F./Gonzalo, J./Peñas, A.: Information Retrieval & Natural Language Processing. Ciudad Universitaria, Dpto. IEEC, Madrid 1999. http://rayuela.ieec.uned.es/~ircourse/

[Witten et al. 1999] Witten, I. et al. : Text Mining: A New Frontier for Lossless Compression. University of Waikato, Computer Science Department, Hamilton (New Zealand) 1999. http://www.cs.waikato.ac.nz/~nzdl/publications/1999/ IHW-ZB-MM-WJT-Text-Mining.pdf

[Wojciech 1999] Wojciech, S.: Partial Parsing for Corpus Annotation and Text Processing. PhD Thesis. University of Saarland, Saarbrücken 1999.

[Yeates 2000] Yeates, S.: Text Mining. University of Waikato, Computer Science Department, Hamilton (New Zealand). 01.09.2000. http://www.cs.waikato.ac.nz/~nzdl/textmining/