

Automatische Indexierung von wirtschaftswissenschaftlichen Texten - ein Experiment

Johann Haller, Bärbel Ripplinger, Dieter Maas



**Institut der Gesellschaft zur Förderung
der Angewandten Informationswissenschaft
an der Universität des Saarlandes**

www.iai.uni-sb.de

Manuela Gastmeyer

Hamburgisches Welt-Wirtschafts-Archiv



Inhalt

1	EINLEITUNG	1
<hr/>		
2	DAS AUTINDEX-SYSTEM	2
<hr/>		
2.1	FUNKTIONSWEISE UND SYSTEMARCHITEKTUR	2
2.2	RESSOURCEN	4
2.2.1	THESAURI	4
2.2.2	KLASSIFIKATIONSSCHEMATA	4
2.2.3	LEXIKA	4
2.3	KOMPONENTEN	4
2.3.1	LINGUISTISCHE ANALYSE MPRO	5
2.3.2	EVALUIERUNG DER TEXTELEMENTE	5
2.3.3	ERMITTLUNG VON WORTGRUPPEN DURCH OBERFLÄCHENPARSING	6
2.3.4	ERGEBNISAUSGABE	6
<hr/>		
3	DAS HWWA-EXPERIMENT	6
<hr/>		
3.1	DIE DOKUMENTE DES HWWA	6
3.2	INDEXIERUNG	7
<hr/>		
4	PROBLEME UND ABHILFEMAßNAHMEN	11
<hr/>		
4.1	PROBLEME 1: ZEICHEN, STRINGS, SPRACHE	11
4.2	PROBLEME 2: NAMEN, FREMDWÖRTER, LÄNDER	12
4.3	PROBLEME 3: THESAURUS, SYNONYME, KOMPOSITA	12
4.4	PROBLEME 4: ALLGEMEIN- VS. FACHWÖRTER	13
4.5	PROBLEME 5: SEMANTIK Ø ZU VIELE DESKRIPTOREN	14
4.6	ABHILFEMAßNAHMEN UND ERSTE FORTSCHRITTE	14
<hr/>		
5	AUSBLICK: FORSCHUNG UND WEITERE EXPERIMENTE	15
<hr/>		
6	REFERENZEN	16
<hr/>		

1 Einleitung

Große Datenbankbetreiber und andere Institutionen, die sich als Wissenspool für ein bestimmtes Spezialgebiet verstehen, benutzen heute zunehmend Softwareprogramme zur Unterstützung des Indexierprozesses. Hierbei stehen dem Indexierer entsprechend aufbereitete Thesauri zur Verfügung, die die Struktur veranschaulichen und helfen, die eventuell existierenden Richtlinien zur Schlagwortvergabe in die Tat umzusetzen. Integrierte Rechtschreibkorrekturprogramme sollen die Eingabe von falsch geschriebenen Schlagworten verhindern, z.B. im Falle von neu vorgeschlagenen Begriffen. Die zeitaufreibende intellektuelle Aufgabe der Ermittlung und Zuweisung der Schlagworte selbst wird jedoch nur in einigen wenigen Forschungsprojekten versuchsweise unterstützt. Die meisten davon arbeiten mit einem kontrollierten Vokabular, oft ein Thesaurus oder eine ‚legalisierte‘ Termliste (z.B. in Salton [7]). Neuere Arbeiten wie der ‚latent semantic indexing approach‘ LSI [3][6] oder der ‚controlled-term approach‘ des Condorcet-Projekts an der Universität Twente [1] führen eine syntaktisch-semantische (Tiefen-) Analyse durch, deren Auswertung auf die Verbindung zu einer Wissensbank angewiesen und deshalb nur für kleine Weltausschnitte operabel ist. Keines dieser Projekte integriert jedoch allgemeine und robuste natürlichsprachliche (NLP-)Analyseverfahren, wie sie in den siebziger Jahren in Projekten wie CONDOR [9] experimentell erprobt worden waren; die Heranführung dieser Projekte an den praktischen Einsatz wurde jedoch durch die noch nicht ausgereifte technische Infrastruktur erheblich erschwert. Ein nur geringer Prozentsatz der Dokumente lag in maschinenlesbarer und verwertbarer Form vor, Hard- und Software für die Bearbeitung großer Datenmengen und für intensive Rechenvorgänge war nur begrenzt verfügbar. Ähnliches galt auch für die Grundlagen und Methoden der NLP (Lexika, Grammatiken, Algorithmen, Formalismen). Heute sind auf all diesen Gebieten große Fortschritte realisiert, und der Versuch scheint berechtigt, dieses komplexe Problem durch die Erprobung geeigneter NLP-Tools erneut einer Lösung zuzuführen.

Das IAI entwickelt seit 1985 ein solches allgemeines Analyseverfahren für Deutsch (und andere europäische Sprachen), das auf den Ressourcen des seit den 70er Jahren bestehenden Sonderforschungsbereichs 100 der Universität Saarbrücken aufbaut. Dieses Analyseverfahren ist inzwischen an einer großen Menge realer Texte aus den verschiedensten Fachgebieten getestet worden und weist eine hohe Effizienz und geringe Fehleranfälligkeit auf. Als eine von mehreren Anwendungen dieses Analysewerkzeugs wird ein Softwarepaket AUTINDEX zur automatischen Extraktion von Schlagworten aus deutschen und englischen Dokumenten entwickelt, das mit linguistischen und statistischen Funktionen arbeitet und die Terminologie eines Thesaurus in die Auswertung einbeziehen kann. Dieses System wird in Kapitel 2 dargestellt.

Die Einsatzmöglichkeiten dieser Software zur Unterstützung bei der Schlagwortvergabe sind an etwa 100 elektronisch aufbereiteten deutschsprachigen Dokumenten aus den Beständen des Hamburgischen Welt-Wirtschafts-Archivs (HWWA) getestet worden, wobei das Vokabular des STANDARD-THESAURUS WIRTSCHAFT als fachsprachliche Basis herangezogen wurde. Die Beispiele stammen aus allen Sammelgebieten des

HWWA: wissenschaftliche Literatur, Branchen- und Produktliteratur, Pressedokumentation.

Das HWWA besitzt eine der weltweit größten Spezialbibliotheken für Literatur aus Wirtschaftswissenschaft und Wirtschaftspraxis sowie angrenzenden Sachgebieten. Anfang 2000 war der Bestand auf rd. 1,1 Mio. Bände und 18 Mio. Presseauschnitte angewachsen. Der jährliche Zugang beläuft sich auf ca. 20.000 Bände und 180.000 Presseauschnitte.

Traditionell werden erhebliche Teile der Bestände dokumentarisch aufbereitet und können themenspezifisch nachgewiesen werden. Ausgewertet werden Monographien, Beiträge aus Sammelwerken, Aufsätze aus Periodika wie Zeitschriften und Jahrbüchern sowie Zeitungsartikel. Gegenwärtig werden etwa 2.000 Zeitschriften und 120 Zeitungen inhaltlich erschlossen. Das HWWA setzt bereits seit Beginn seines Bestehens eigene Erschließungsmedien ein, die kontinuierlich weiterentwickelt worden sind. Gegenwärtig bildet das Vokabular des STANDARD-THESAURUS WIRTSCHAFT (STW) die Basis für die Inhaltsauswertung.

Durch die stetige Zunahme vernetzter und virtueller Informationsangebote unterschiedlichster Herkunft über Bibliotheksverbände und Internet spielt eine differenzierte Strukturierung und Organisation der Wissensbestände und –nachweise eine größere Rolle denn je, auch für Einrichtungen wie das HWWA. Die erweiterten technischen Funktionalitäten lassen die Anforderungen an einen wissenschaftlichen Informationsservice ständig steigen, während die finanziellen und personellen Ressourcen reduziert werden. Daraus folgt, dass zukünftig nur dann ein hochwertiges Dienstleistungsangebot aufrechterhalten und durch den Nachweis elektronischer Informationsquellen ausgebaut werden kann, wenn es gelingt, einen Teil der bislang ausschließlich intellektuell erbrachten Sacherschließungsleistungen durch automatisierte Verfahren zu substituieren und zu ergänzen. Vor diesem Hintergrund wurde die Gelegenheit genutzt, gemeinsam mit dem IAI ein Projekt zur automatischen Sacherschließung durchzuführen.

Über dieses Experiment wird in Kapitel 3 berichtet.

2 Das AUTINDEX-System

2.1 Funktionsweise und Systemarchitektur

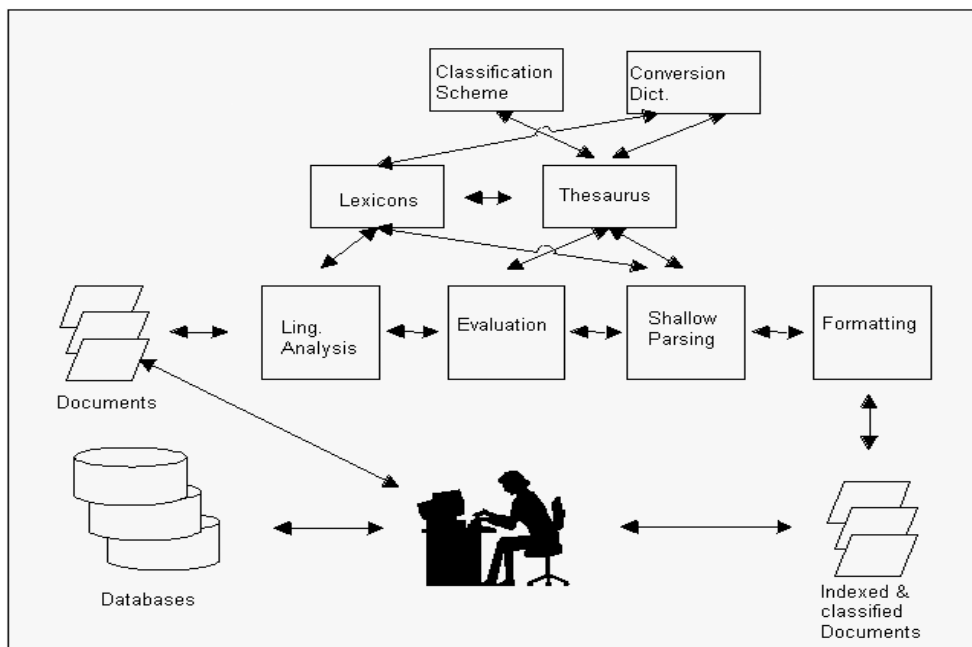
AUTINDEX benützt als Grundlage umfangreiche linguistische Werkzeuge, wobei das Kernsystem aus einem morpho-syntaktischen Analyseverfahren (Mpro) besteht. Diese Analyse weist jedem Wort im Dokument neben grammatikalischen Informationen wie der Wortklasse auch semantische Merkmale zu. Für deutsche Texte wird eine Komposita-Analyse durchgeführt, so dass auch Informationen über die möglichen Bestandteile eines Wortes verfügbar sind. In einem zweiten Schritt werden die bedeutungstragenden Elemente (Grundformen von Nomen und Verben) gesammelt und aus den zugewiesenen semantischen Informationen die im Dokument am häufigsten auftretenden semantischen Klassen ermittelt. Danach werden alle Elemente, die diesen Klassen zugeordnet sind, gesammelt und gewichtet. Das Gewicht des Wortes ist dabei eine Funktion der Frequenz und dem Gewicht, das der jeweiligen semantischen Klasse

zugeordnet ist. Auch die Bestandteile von Komposita werden in den Gewichtungsalgorithmus mit einbezogen.

In einem dritten Schritt wird zusätzlich ein sog. ‚Shallow Parsing‘ durchgeführt, dessen Ergebnisse die Grundlage der Erkennung von Mehrwortlexemen und deren syntaktischer Varianten sind. Mehrwortlexeme, die den in Schritt 2 ermittelten semantischen Klassen zugeordnet werden können, bilden zusammen mit den Einwort-Termen die Menge der sog. Schlüsselwörter. Wenn die Indexierung einen vorhandenen Thesaurus als Basis benützt, werden die Schlüsselwörter mit den Elementen des entsprechenden Thesaurus abgeglichen. Das Ergebnis dieser Operation ist die Liste der sog. Deskriptoren. Die Zahl der einem Dokument zugewiesenen Deskriptoren kann über einen Parameter vom Benutzer bestimmt werden. Zu diesen Deskriptoren werden zusätzlich noch die Oberbegriffe bestimmt und extra ausgegeben.

Zur Klassifikation eines Dokuments wird die Information statistisch ausgewertet, die bestimmten Begriffen im Lexikon oder im Thesaurus zugewiesen ist. Die am häufigsten vorkommenden Klassen werden dem Dokument als Klassifikation zugewiesen, wobei auch hier der Benutzer die Anzahl der zugewiesenen Klassen steuern kann. Existiert kein anwendungsspezifischer Thesaurus, wird der ebenfalls im Lexikon abgelegte NACE-Code (Wirtschaftsbranchen in der EU) zugrunde gelegt.

Die folgende Abbildung zeigt, wie die Einzelkomponenten des AUTINDEX-Systems interagieren, und welche Ressourcen jeweils benutzt werden. Die Rolle des menschlichen Bearbeiters besteht in diesem Szenario darin, die Dokumente für den Indexierungsprozess auszuwählen sowie das Ergebnis der automatischen Indexierung zu evaluieren, d.h. zu validieren oder zu verwerfen. Nur der Titel wird als bibliographische Information benutzt, soweit er besonders gekennzeichnet ist.



Das Gesamtsystem wurde bereits einem umfangreichen Test mit Texten und dem Thesaurus des FIZ Technik unterworfen, in dessen Verlauf die Effizienz und Genauigkeit erheblich gesteigert werden konnte.

2.2 Ressourcen

2.2.1 Thesauri

Eine spezielle Schnittstelle steht zur Anbindung eines benutzerdefinierten Thesaurus zur Verfügung; dies geschieht in Form einer Konvertierung und einer linguistischen Analyse der Thesauruseinträge nach dem gleichen Prinzip.

2.2.2 Klassifikationsschemata

Ein Standardschema (NACE- und Ländercode) steht zur Verfügung, kann jedoch vom Benutzer angepasst werden

2.2.3 Lexika

Die sprachbezogenen Lexika enthalten morphologische, syntaktische und semantische Informationen sowie Hinweise auf die Zugehörigkeit zu Fachgebieten, die besonders bei bilingualer Verwendung für die Übersetzung benutzt werden. Das deutsche Lexikon enthält derzeit 60.000 Stämme (gleichbedeutend mit ca. 150.000 Grundformen und Ableitungen, ca. 6 Millionen flektierte Wortformen, unbegrenzte Kompositaerkennung), das englische Lexikon 40.000 Stämme (mit wesentlich weniger Ableitungen).

Als Beispiel für Lexikoneinträge sollen die beiden Elemente dienen, die für die Analyse des Wortes Organisationsberatung benötigt werden:

```
{string=rat,c=v,fuge=e,prspara=et,ptcpara=en,ptzge=ge,  
n={er=agent&anto,ung=massnahme},a={bar=able,sam=va},  
novzs=durcheinander;aneinander,lu=raten}
```

```
{string=organis,c=v,n={ator=agent&anto,ation=coll;massnahme},  
v={ier=nc},nopr=er,lu=organisieren}
```

Neben der syntaktischen Kategorie wird im Lexikon vor allem festgehalten, mit welchen Morphemen das Partizip des entsprechenden Verbs gebildet wird (prspara, ptcpara und ptzge), welche nominalen Ableitungen gebildet werden können und welche semantischen Interpretationen diesen Ableitungen zugeordnet werden können, sowie einige explizite nicht zugelassene Verbindungen, die zur Beschleunigung der Analyse hilfreich sind.

2.3 Komponenten

AUTINDEX gliedert sich in vier Programmteile, die sequentiell hintereinander ablaufen: linguistische Analyse, Evaluierung der Textelemente, Wortgruppenermittlung (Oberflächenparsing), Ergebnisausgabe.

2.3.1 Linguistische Analyse Mpro

Das Kernsystem der linguistischen Analyse besteht aus Lexikon und Grammatik. Das Lexikon ist eine Menge von Einträgen, die als Merkmalbündel kodiert sind. Jeder Lexikoneintrag beinhaltet die Angabe eines Strings, der das jeweilige Allomorph darstellt. Diese Strings werden bei der Zerlegung der Wortformen verwendet.

Jede Wortform ist nach Beendigung der Analyse mit zwei Informationsblocks verknüpft:

- Merkmalbündel für die flexionsmorphologischen Features
- Wortstruktur (Derivation und Komposition)

Bei Substantiven und Verben werden die üblichen morpho-syntaktischen Merkmale ermittelt, bei Adjektiven wird nur die Endung (e,es,em,en,er) festgehalten, da deren Informationsgehalt je nach den Wünschen der Anwender dargestellt wird.

Die Wortstruktur ist eine syntaktisch-semantische Interpretation der gegebenen Wortform, wenn es sich um ein komplexes Wort handelt. Sie ist identisch mit der Zitatform des Wortes, wenn das Wort einfach ist. In jedem Fall wird auch die Zitatform berechnet ("Lemma").

Die Grundidee von Mpro ist, dass sich komplexe Wörter aus einfachen ableiten lassen, so dass ein komplexes Wort automatisch mit einer Paraphrase aus seinen Bestandteilen erklärt werden kann. Das lexikalische Material für das Deutsche (jetzt etwa 50.000 Morpheme, die von etwa 60.000 Allomorphen repräsentiert werden) stammt nicht aus handelsüblichen Lexika, sondern aus Texten (Zeitschriften, medizinischen Diagnosen, Reparaturmanualen, Gelben Seiten, Arbeitsberichten, Krimis, verschrifteten Ansprachen usw.). Es ist dadurch wesentlich realitätsnäher als etwa in Lexika über einen langen Zeitraum gesammeltes Material, das oft zum großen Teil aus literarischen Quellen stammt.

Im folgenden zwei Beispiele für die morphologische Analyse von Komposita:

```
{ori=Organisationsberatung,c=noun,lu=organisationsberatung,  
ls=organisieren#be$raten,ss=coll;massnahme#massnahme,nb=sg,  
g=f}
```

```
{ori=Fischfang,c=noun,lu=fischfang,ls=fisch#fangen,  
ss=animal&ed#act;result,case=nom;dat;acc,nb=sg,g=m}
```

Die durch Semikolon getrennten semantischen Merkmalsalternativen spiegeln die Tatsache wider, dass ‚Fang‘ sowohl die Tätigkeit als auch das Ergebnis des ‚Fangens‘ sein kann; ähnliches gilt für ‚Organisation‘ als Tätigkeit und kollektive Gruppierung.

2.3.2 Evaluierung der Textelemente

In dieser Phase werden die Textelemente, die zu den Wortklassen Nomen, Verb und Adjektiv gehören (den sogenannten bedeutungstragenden Wortklassen), daraufhin analysiert, zu welchen semantischen Klassen sie mehrheitlich gehören. Wörter (oder Wortteile von Komposita – bisher Head oder Komplement ohne Differenzierung), deren semantische Klasse z.B. nur ein einziges Mal im Text vorkommt (z.B. ‚Erosion‘ als Spezialbegriff der Geographie bei „Erosion der Steuergesetze“), werden dann von der weiteren Betrachtung ausgeschlossen.

Es werden nur die häufigsten semantischen Klassen für die Weiterverarbeitung benutzt, deren Anzahl der Benutzer festlegen kann.

2.3.3 Ermittlung von Wortgruppen durch Oberflächenparsing

Hierzu muss eine Bestimmung der Strukturierung des Satzes in Teilgruppen vorgenommen werden, ein sogenanntes „Oberflächenparsing“

Mpro verwendet hier eine externe Grammatik, die die Erkennung syntaktischer Strukturen einer bestimmten Sprache, hier des Deutschen, beschreibt. Diese Grammatik ist in Teilgrammatiken gegliedert, die sequentiell ablaufende Regeln mit Kontextspezifikationen enthalten. Durch diese Strategie entsteht ein robustes Verfahren, das zwar durch die Nichtverfolgung von Mehrdeutigkeiten eine gewisse Unschärfe in Kauf nimmt, jedoch in der überwiegenden Anzahl der Fälle Ergebnisse liefert, die auch bei partieller Analyse recht gut brauchbar sind.

2.3.4 Ergebnisausgabe

Die in derselben Weise analysierten Einträge des Thesaurus werden dann mit ihren Entsprechungen in den Texten in Verbindung gebracht und ggf. als Mehrwortbegriffe vorgeschlagen. Im Deutschen dient dieses Verfahren meist nur dazu, im Thesaurus enthaltene Komposita auch dann an einen Text zu binden, wenn die Teile im Text getrennt vorhanden sind (z.B. als Präpositionalphrase). In anderen Sprachen, z.B. Englisch, sind auch die Thesauruseinträge meist Mehrwortbegriffe.

Als Ausgabe werden dann folgende Elemente präsentiert (Beispiele in der Darstellung des Experiments):

- Eine Liste der wichtigsten Wörter des Textes (mit ihrem errechneten Gewicht)
- Eine Liste der Deskriptoren (vom Thesaurus erlaubte Begriffe)
- Eine Liste der nicht im Thesaurus enthaltenen Schlüsselwörter (als Vorschlagsliste)
- Der Ländercode
- Der NACE-Branchencode (wenn gewünscht)
- Aus dem Text ermittelte Namen von Personen und Firmen

3 Das HWWA-Experiment

3.1 Die Dokumente des HWWA

Die 89 deutschsprachigen Dokumente des HWWA wurden aus verschiedenen Gebieten ausgewählt, die für die Sammeltätigkeit der Institution charakteristisch sind: wissenschaftliche Literatur, Branchen- und Produktliteratur sowie der Pressedokumentation.

Die Texte lagen in verschiedenen Formaten vor: entweder schon maschinenlesbar im PDF-Format, oder in einer gescannten und nach ASCII konvertierten Form. Hierbei wurden keine Strukturierungen des Textes erhalten, Hinweise auf den Platz der Überschrift, der Autoren oder eventueller Stichwörter gab es nicht. Rechtschreibfehler

wurden soweit wie möglich korrigiert. In dieser Form wurden die Texte unmittelbar der linguistischen Analyse unterworfen.

Der Thesaurus wurde ebenso in einer ASCII-Version übernommen und in gleicher Weise linguistisch analysiert.

3.2 Indexierung

Hieran schlossen sich insgesamt 4 Indexierungen an, 2 durch menschliche Indexierer, die auf Anweisung genauer und ausführlicher als sonst indexierten, z.B. mehr Schlagworte vergaben als die sonst üblichen 2-3 pro Dokument (bezeichnet als I1 bzw. I2). Die beiden anderen Indexate wurden mit zwei Versionen der automatischen Indexierung erstellt: zunächst mit einer älteren (mehr allgemeinsprachlich orientierten) Version (MULTIDESC), die für die Indexierung und Klassifikation von Wirtschaftsnachrichten getestet wurde, sowie der Version INDEXIERUNG. Diese Version war mit Abstracts im technischen Bereich entwickelt und bereits einmal mit dem Thesaurus des Fachinformationszentrums Technik getestet worden. Eine Erweiterung erfolgte hier für längere Texte (wie die des HWWA), denen auch mehr Deskriptoren zugewiesen wurden.

Bei MULTIDESC wurden nur jeweils 5 semantische Klassen in die Indexierung einbezogen, die in diesem Text besonders häufig vertreten waren, während bei INDEXIERUNG alle semantischen Klassen einbezogen wurden; außerdem wurden in MULTIDESC die Teile von Komposita gleich behandelt, während in INDEXIERUNG eine Differenzierung der Haupt- und Nebenelemente vorgenommen wurde.

Die Ergebnisse dieser Indexierungen wurden dann daraufhin verglichen, inwieweit die fortgeschrittenere INDEXIERUNG mindestens die auch von den menschlichen Bearbeitern vergebenen Schlagworte ermittelten. Hierbei ergaben sich folgende statistischen Werte:

Statistik

	Anzahl der Deskriptoren:		Prozentergebnisse: durch Indexierung gefundene Deskriptoren
	In dem intellektuellen Indexat	davon durch Indexierung ermittelt	
Indexat 1 (I1)	658	344 84*	52,3 Prozent
Indexat 2 (I2)	626	309 72*	49,3 Prozent

* Anzahl der intellektuellen Deskriptoren, die durch das Programm unter "Oberbegriffe" erscheinen

An den folgenden zwei Textbeispielen soll dieser Vorgang verdeutlicht werden; außerdem wird auch sichtbar, wie der Weg von den linguistischen Begriffen zu den Deskriptorkandidaten aussieht

Beispiel 1: Organisationale Veränderung

Michael Meyer / Peter Heimerl-Wagner

Organisationale Veränderung: Transformationsreife und Umweltdruck

Intervention; Organisationsberatung; Struktureller Wandel; System-Umwelt-Kopplung

Der Wandel und die Veränderung von Organisationen ist eines der zentralen Themen der modernen Organisationsforschung. Aus unterschiedlichen theoretischen Perspektiven - Organisationsentwicklung, geplanter Wandel, lernende Organisation rücken Veränderungsprozesse ins Blickfeld der wissenschaftlich und praktisch interessierten Beobachter. In jüngerer Zeit dominieren dabei Konzepte wie Lean Management, Business Reengineering, Total Quality Management und Learning Organization die Diskussion, wobei meist Bestseller der populärwissenschaftlichen Managementliteratur (z.B. ...) die Initialzündung gaben. Dabei wird regelmäßig mit einer positiven Konnotation des Veränderungsbegriffes gearbeitet und nicht hinterfragt, ob Veränderung für Organisationen überhaupt wünschenswert ist.

Für diesen Text wurden die folgenden **Indexate** ermittelt:

I1: Organisatorischer Wandel, Organisationsforschung, Systemtheorie, Theorie \emptyset vom Programm ebenfalls alle ermittelt

I2: Organisationsberatung, Organisatorischer Wandel, Organisationstheorie, Systemtheorie, Theorie \emptyset vom Programm ebenfalls ermittelt

I2 zusätzlich: Organisationssoziologie \emptyset vom Programm nicht ermittelt

Das ausführliche Ergebnis der automatischen Indexierung ist – wie im ersten Teil beschrieben – in mehrere Abschnitte eingeteilt:

Schlüsselwörter sind Wörter, die die linguistische Analyse als wichtig beurteilt und mit einer entsprechenden Punktwertung versehen hat:

Organisation[63]; Veränderung[35]; Management[15]; Theorie[8];
Organisieren[7]; Organisationsforschung[6]; Entscheidung[5];
Entwicklung[3]; Intervention[2]; Dynamisierung[2]; Lernende
Organisation[2]; Literatur[2]; Strategie[2]; Beobachtung[2];
Analyse[2]; Wandel[1]; Umwelt[1]; Systemtheorie[1]; Diskussion[1];
Interventionsstrategie[1]; Einsatz[1]; Flexibilisierung[1];
Auflösung[1]; Klassifikation[1]; Organisationsanalyse[1];
Organisationsberater[1]; Organisationsberatung[1];
Organisationsgrenze[1]; Organisationsintern[1]; Beratung[1];
Reformulierung[1]; Selbstorganisation[1]; Stabilisierung[1];
Steuerung[1]; Differenzierung[1]; Thema[1]; Kontingenztheorie[1]

Deskriptoren sind Begriffe, die mit einer Wertigkeit > 0 versehen wurden und im Thesaurus enthalten sind:

Organisation; Management; Theorie; Organisationsforschung;
Entscheidung; Entwicklung; Intervention; Lernende Organisation;
Literatur; Strategie; Umwelt; Systemtheorie; Klassifikation;
Organisationsberatung; Selbstorganisation; Kontingenztheorie

Sie sind nur dann nicht in die Menge der Deskriptoren aufgenommen worden, wenn ein spezifischerer Unterbegriff selbst in dieser Liste steht.

Eine Zusammenfassung der übriggebliebenen Schlüsselwörter zeigt relativ rasch, dass es sich hier in den meisten Fällen um ‚Allgemeinwörter‘ handelt, die grundsätzlich bereits bei der linguistischen Indexierung herausgenommen werden könnten, wenn man sie nicht wie die beiden am höchsten gewichteten Begriffe ‚Veränderung‘ und ‚Organisieren‘ in eine spezielle Spalte ‚allgemeine Prozesse‘ einordnen möchte. Ein linguistisches Indiz hierfür wäre auch die mehreren Begriffen gemeinsame Eigenschaft der Ableitung von Verben:

Nicht-Deskriptoren

Veränderung[35]; Organisieren[7]; Dynamisierung[2];
Beobachtung[2]; Analyse[2]; Wandel[1]; Diskussion[1];
Interventionsstrategie[1]; Einsatz[1]; Flexibilisierung[1];
Auflösung[1]; Organisationsanalyse[1]; Organisationsberater[1];
Organisationsgrenze[1]; Organisationsintern[1]; Beratung[1];
Reformulierung[1]; Stabilisierung[1]; Steuerung[1];
Differenzierung[1]; Thema[1]

Spezialbegriffe (z.B. Komposita wie ‚Interventionsstrategie‘) könnten in eine Vorschlagsliste für neu in den Thesaurus aufzunehmende Stichwörter eingespielt werden.

Von einem anderen Experiment mit Indexierung lieferte das Programm auch eine ‚Spartenzuordnung‘, die angibt, zu welchem Produktionszweig das Dokument gehören kann:

Sparte

N7300 Forschung und Entwicklung[72]; N7410 Rechts-, Steuer- und
Unternehmensberatung, Markt- und Meinungsforschung,
Beteiligungsgesellschaften[68]; N7414 Unternehmens- und Public-
Relations-Beratung[68]

Aus verschiedenen Gründen waren dem Programm auch einige Textstrings unbekannt:

Unbekannt

Lean ; Reengineering ; Quality ; Learning ; Organization ; i ; change ;
agents ; autopoietisch ; Luhmann ; open ; approach ;
Selbstreferentialität ; organizational ; selegieren

Hierbei handelte es sich in der Hauptsache um englische Begriffe, die nach den Angaben des HWWA allerdings gehäuft in wirtschaftswissenschaftlichen Artikeln auftauchen. Die ‚eingedeutschten‘ Fremdwörter ‚Autopoietisch, ‚Selbstreferentialität‘ und ‚selegieren‘ können ins Wörterbuch aufgenommen werden. Der Name ‚Luhmann‘ kann inzwischen durch eine spezielle Heuristik der Namenerkennung als solcher klassifiziert werden.

Beispiel 2: Dänemarks Fischwirtschaft im Umbruch

Fangquoten gehen zurück

Aquakultur wird wichtiger

Nachdem die Europäische Union die dänischen Fangquoten drastisch gesenkt hat, haben sich die Zukunftsprognosen für die Fischerei des weltweit viertgrößten Exporteurs von Fischen und Fischprodukten verschlechtert. Hoffnung setzt die Branche auf die Aquakultur, die in Dänemark insbesondere mit der Zucht von Aalen beachtliches Wachstum erzielt.

Wenn man von Jütland einmal absieht, besteht Dänemark nur aus Inseln - 406 sind es insgesamt. Kein Bewohner dieses kleinen skandinavischen Landes wohnt mehr als 52 Kilometer vom Meer entfernt. Kann es bei soviel Nähe zur See verwundern, dass die dänische Fischwirtschaft - zumindest gemessen an der Fangmenge - die größte Europas ist? Entsprechend machen Fisch und Fischprodukte 4 Prozent der dänischen Exporte aus.

Indexate

I1: Dänemark, Hafen, Aquakultur, Fischwirtschaft, Fisch, Fischerei ∅ vom Programm alle ermittelt

I2: Dänemark, Aquakultur, Fischwirtschaft, Fischerei, Produktion, Export ∅ vom Programm ermittelt

I2 zusätzlich: Hafenumschlag ∅ vom Programm nicht ermittelt (dafür ‚Hafen‘!)

Die wichtigsten unter den folgenden Schlüsselwörtern sind wieder mindestens einschlägig; hier wäre durch eine einfache Übernahme einer bestimmten Zahl der (ihrer Wichtigkeit nach geordneten) Begriffe eine relativ gute Abdeckung zu erreichen, wenn diese dann auch nicht mehr in der Deskriptorenliste auftauchen:

Schlüsselwörter

Süßwasserfisch[23]; Produktion[15]; Fischprodukt[10]; Nordsee[10];
Export[6]; Hochseefischerei[5]; See[3]; Sektor[3]; Konsumfisch[3];
Land[3]; Fischart[3]; Regenbogenforelle[3]; Fischindustrie[2];
Norden[2]; Osten[2]; Krebs[2]; Bodenfisch[2]; Branche[2];
Konsumfischanlandung[2]; Konsum[1]; Steinbutt[1]; Forellenei[1];
Sprotte[1]; Fischbestand[1]; Plattfische[1]; Prozent[1]; Industrie[1];
Kattegat[1]; Meerwasserfarm[1]; Skagerrak[1]; Ostsee[1]; Anstieg[1];
Auktion[1]; Meeresfischfarm[1]; Produktionskosten[1]

Deskriptoren

Dänemark; Süßwasserfisch; Produktion; Fischprodukt; Nordsee;
Export; Hochseefischerei; See; Fischindustrie; Branche; Osteuropa;
Frankreich; Asien; Schweden; Konsum; Industrie; Ostsee; Auktion;
Produktionskosten

Schwieriger ist es hier, Kandidaten für eine Ergänzung des Thesaurus automatisch hervorzuheben, und auch insbesondere die Länder auszuschneiden, die als kurz genannte Zielländer für den Export dänischer Fische für das Kernthema des Aufsatzes nur von geringer Wichtigkeit sind:

Nicht-Deskriptoren

Sektor[3]; Konsumfisch[3]; Land[3]; Fischart[3];
Regenbogenforelle[3]; Norden[2]; Osten[2]; Krebs[2]; Bodenfisch[2];
Konsumfischanlandung[2]; Steinbutt[1]; Forellenei[1]; Sprotte[1];
Fischbestand[1]; Plattfische[1]; Prozent[1]; Kattegat[1];
Meerwasserfarm[1]; Skagerrak[1]; Anstieg[1]; Meeresfischfarm[1]

Leicht konnte das Programm hier die Sparte bestimmen:

Sparte

N0500 Fischerei und Fischzucht[100]; N1520 Fischverarbeitung[31];
N5138 Großhandel mit sonstigen Nahrungsmitteln[18]; N5223
Einzelhandel mit Fisch und Fischerzeugnissen[18]

Bei der speziellen Ermittlung der Länder stand Dänemark mit so großem Abstand vor allen anderen Ländern, dass es hier als einziges betroffenes Land genannt wurde:

Länder

Dänemark C4EUDE[100]

Zu den unbekanntem Begriffen gehören hier neben kleineren dänischen Städten und Häfen noch (die auch Word 2000 unbekannt) Limande sowie der Stintdorsch. Ein nicht entdeckter Rechtschreibfehler aus dem Scansvorgang (Apuakultur) ergänzt die Liste:

Unbekannt

Apuakultur ; Isefjord ; Limfjord ; Belten ; a ; i ; Limande ; Stintdorsch
; Wittling ; Esbjerg ; Thyboren ; Skagen ; Lolland ; Falster ; Kabayaki

4 Probleme und Abhilfemaßnahmen

Generell können die aufgetauchten Probleme in drei Kategorien eingeteilt werden: kurzfristige Abhilfemaßnahmen, deren Strategie nicht weiterer Überlegung bedarf, mittelfristige Arbeiten, die einer ausführlicheren Konzeption bedürfen, und Aktivitäten, die eher der längerfristigen Forschung zuzurechnen sind.

4.1 Probleme 1: Zeichen, Strings, Sprache

Der gesamte HWWA-Test wurde (wie in Kap. 3.1. beschrieben) als ‚realer‘, d.h. so praxisnaher Test wie möglich durchgeführt: es wurde an den Dokumenten keinerlei Vorbearbeitung unternommen, die auf die Indexierung ausgerichtet war.

So blieben diese z.T. völlig unstrukturiert, es wurde keine Unterscheidung zwischen Überschrift, Quellenangabe, Autorenherkunft, Publikationsangaben und dem Text selbst eingeführt. Dies hatte natürlich zur Folge, dass in HWWA-Dokumenten die Stadt Hamburg als Deskriptor auftauchen konnte, obwohl im Dokument nicht die Rede davon war.

Ebenso ermittelte beispielsweise das Scan-Programm unterschiedliche Arten von Bindestrichen, die dann auch unterschiedlich interpretiert werden konnten; dies führte zur Nichterkennung von nach Silben getrennten Wörtern.

Auch die Rechtschreibung wurde nur mit einem gängigen Korrekturprogramm verbessert, das Alte und Neue Rechtschreibung gleichzeitig akzeptierte; so konnte ein Deskriptor ‚Weißrussland‘ nicht ermittelt werden, wenn er als ‚Weißrußland‘ noch im Thesaurus stand. Eine konsequente Umstellung von Thesaurus und Dokumenten auf die Neue Rechtschreibung wird hier Abhilfe schaffen.

Ähnliche Probleme treten bei der Schreibung von neu geschaffenen Begriffen auf, z.B. dem eCommerce resp. E-Commerce; hier kann per Programm ein gewisses Fuzzy-Matching erreicht werden, das jedoch enge Grenzen hat, wenn damit wieder zu viele ähnliche Wörter mit einbezogen werden.

Viele der Texte enthielten auch Teile in englischer Sprache, manche einzelne Begriffe, andere ganze Abstracts in Englisch. Hier kam es manchmal zu zufälligen Übereinstimmungen zwischen den Elementen beider Sprachen; so kann englisch ‚asset‘ von der deutschen Analyse als Vergangenheit 2. Person Plural von ‚essen‘ analysiert werden, und ‚Capital‘ als differierende Schreibweise des deutschen ‚Kapital‘. Eine systematische Abgrenzung bzw. Markierung fremdsprachlicher Textteile oder eine Gesamtformatierung der Texte in SGML würde dieses Problem vermeiden.

4.2 Probleme 2: Namen, Fremdwörter, Länder

Eine zweite Gruppe von noch relativ einfach zu behebbenden Problemen ergibt sich durch das Auftreten von Eigennamen (Personen und Länder) sowie von einzelnen Begriffen in anderen Sprachen. Wenn ein dem Lexikon nicht bekannter Eigenname zerlegt werden kann, ergeben sich zufällige Schlüsselwörter (etwa ‚Ei‘ aus ‚Eizenstat‘ oder ‚Mann‘ aus ‚Mannesmann‘ bzw. ‚Fisch‘ aus (Außenminister) Fischer); dies kann durch den Einbau einer automatischen Namenerkennung weitgehend reduziert werden, die mit bestimmten linguistischen Patterns (Herr, Frau, Minister und andere Titel) arbeitet.

Länder sollten eine gesonderte Behandlung erfahren, da hier auch detaillierte Indexrichtlinien des HWWA vorliegen (die beim hier beschriebenen ersten Versuch noch nicht zugrundelagen). Länder sollen auch indexiert werden, wenn z.B. die entsprechende Hauptstadt genannt ist, andererseits sollen mehrere Länder unter bestimmten Umständen zu ‚Europa‘ oder gar ‚Welt‘ zusammengefasst werden.

Etwas schwieriger ist die Bearbeitung von mehrgliedrigen englischen Begriffen, die oftmals schwer als zusammengehörig erkannt werden können, wenn sie mitten in einem deutschen Satz stehen; vor allem können hier keine Varianten geprüft werden. Trotzdem kann ein Teil dieser Ausdrücke (‚Corporate Governance‘, ‚Conjoint-Analysis‘) mit einfachen programmtechnischen Maßnahmen mit einbezogen werden.

4.3 Probleme 3: Thesaurus, Synonyme, Komposita

Zu diesem Problemkreis gehören alle fehlerhaften Deskriptoren, die durch eine „zu exakte“ Interpretation des Thesaurus und seiner Relationen verursacht wurden. So ist im Thesaurus eine Relation enthalten, die den Begriff „Wind“ durch den Begriff „Sturm“

ersetzen lässt. Dies ist für einen Artikel über Windenergie allerdings nicht angebracht – hier sollte der Begriff „Wind“ erhalten bleiben oder nur innerhalb der Zusammensetzung als Indexat herangezogen werden.

In anderen Fällen wurden von den menschlichen Indexierern auch Begriffe eingetragen, die (vermutlich wegen ihrer Neuheit) noch nicht im Thesaurus selbst stehen (Beispiel „Anleihemarkt“); bei anderen Beispielen sind leicht unterschiedliche Varianten verzeichnet, so etwa im Falle der „regenerativen Energie“, die als „erneuerbare Energie“ im Text auftritt, wobei bei anderen Deskriptoren durchaus auch das Adjektiv „erneuerbar“ im Thesaurus steht. Ähnliches ergibt sich bei der Synonymie zwischen (Fließ-)„Fertigung“ und (Fließ-)„Produktion“.

Wie zu sehen, ergeben sich bei einem solchen Experiment auch wertvolle Erkenntnisse für die Gestaltung des Thesaurus selbst, der für die Anwendung eines maschinellen Verfahrens natürlich viel genauer kontrolliert werden muss als für einen menschlichen Benutzer.

Sowohl für die Gestaltung des Thesaurus als auch für die Entscheidung, ob ein Begriff des Textes Deskriptor werden kann, müssen Grundprinzipien für die Aufstellung und Interpretation von Komposita ermittelt und eingehalten werden, sonst ist nicht zu verhindern, dass aus „Bodenerosion“ auch „Boden“ und aus „Werbewirkungsforschung“ auch „Wirkung“ und „Forschung“ als Deskriptoren extrahiert werden, wenn diese als Thesauruseinträge existieren. Einige dieser Begriffe sind durch ihre Einstufung als „allgemeine“ Wörter eventuell auszuschalten – siehe den folgenden Abschnitt, ebenso wie die Bestandteile von Mehrwortbegriffen wie „Neuer Markt“, aus dem man dann auch nicht den „Markt“ alleine als Deskriptor für dieses Dokument haben will.

4.4 Probleme 4: Allgemein- vs. Fachwörter

Begriffe, die auch häufig in der Allgemeinsprache verwendet werden, wie „Messung“, „Ablauf“, „Bewertung“ und ähnliche, meist nicht zusammengesetzte Wörter, werden in den „Sacherschließungsregeln“ des HWWA in eine besondere Kategorie eingeordnet: sie sollen nur zur Ergänzung oder Spezifizierung vergeben werden, auch innerhalb einer sogenannten „Deskriptorenkette“. Für die automatische Indexierung bedeutet dies zunächst, dass ihre Gewichtung sehr niedrig angesetzt werden kann.

Schwieriger ist das Problem der Begriffe zu lösen, die in der Allgemeinsprache und in der Fachsprache eines bestimmten Fachgebietes vorkommen; so hat etwa „Stadt“ in der Regional- und Städteplanung, „Gesellschaft“ in der Soziologie und „Management“ in der Betriebswirtschaft spezielle Interpretationen. In vielen Dokumenten anderer Fachgebiete können sie trotzdem häufig allein und auch in Komposita auftreten. Dieses Problem ist nur durch eine automatische Klassifikation und Einordnung von Dokumenten in ein bestimmtes Fachgebiet zu bewältigen; hierbei müssen jedoch statistische Verfahren angewandt werden, die nicht nur die Häufigkeit eines Begriffs und seines semantischen Feldes im Dokument, sondern auch die relative Häufigkeit von Begriffen in einer gesamten Dokumentmenge und das entsprechende Clustering berücksichtigen.

Ähnlich gelagert ist das Problem der sprachlichen Metaphern: wenn ein Dokument die „Erosion der Steuergesetze“ behandelt, will man nicht die „Erosion“ aus der Geologie

als Deskriptor bestimmen. Zum Teil ist dies über die Frequenz solcher Begriffe lösbar; allerdings ist auch die durchgängige Verwendung z.B. der Wasser- und Meeresmetapher aus Wirtschaftstexten bekannt, wo bekanntlich Geld ebenso fließen und Ströme bilden kann.

4.5 Probleme 5: Semantik Ø zu viele Deskriptoren

Die im präsentierten Verfahren verwendete Methode, die den Wörtern zugeordneten semantischen Felder zur Berechnung des Deskriptorengewichts heranzuziehen, führt in einzelnen Fällen zur Einbeziehung von Deskriptoren, die zwar logisch durchaus in Frage kämen, aber bei genauer Betrachtung des Textes vermutlich eher nicht herangezogen würden. So wird bei einem Text zum „Volkseinkommen“ das Wort „Strafe“ (trotz geringer Frequenz) mit einem entsprechenden Gewicht belegt, da es als semantisches Merkmal das im ganzen Text sehr wichtige „Geld“ aufweist – und im Text ist es tatsächlich als eine von vielen Möglichkeiten für die Ausgaben eines durchschnittlichen Haushalts erwähnt. Sogar der Thesaurus weist (allerdings unter dem Oberbegriff „Justiz“ eine Beziehung zu „Geldstrafe“ auf!). Ein ähnliches Beispiel ist „Benzin“ in einem Text über Drogen: wie diese ist Benzin ein chemischer Stoff (der in manchen Ländern auch als Inhalierdroge benützt wird), im Text wird er allerdings nur als Treibstoff für die Transportautos der Drogendealer erwähnt.

Gerade die letzten Beispiele zeigen auch die Grenzen der automatischen Indexierung auf – die, wie bereits erwähnt, ja nur einen Vorschlag für den menschlichen Indexierer liefern soll, den dieser dann bestätigen oder verwerfen kann.

4.6 Abhilfemaßnahmen und erste Fortschritte

Als kurzfristige Verbesserungsmaßnahmen können also folgende Stichworte genannt werden:

- neue Sonderzeichenbehandlung (Bindestriche)
- Komposita- und Multi-Word-Behandlung
- Namenerkennung
- Englische Mehrwortbegriffe
- Allgemeinwörter in getrennter Liste halten
- Fachwörter nur bei Fachtext verwenden
- Begrenzung der Deskriptorenzahl (nur wichtigste semantische Klassen)
- Überschriften/Anfang gewichten (wenn erkennbar)

Mit diesen kurzfristigen Maßnahmen, die bereits während der Abfassung dieses Berichts in Angriff genommen wurden, soll das Ziel einer Ermittlung von ca.75% der intellektuellen Deskriptoren erreicht werden, was zu einem ersten Probeeinsatz der automatischen Indexierung ausreichen dürfte.

So konnte durch die ersten beiden Schritte der Prozentsatz der gefundenen Deskriptoren bereits auf 60% erhöht werden: komplexe Deskriptoren wie „betriebliches Informationssystem“, „Organisationsstruktur“, „Öffentlicher Bau“, „Verkaufsraumgestaltung“, „Werkzeugindustrie“, „Privater Haushalt“ wurden mit ermittelt, die vorher gefehlt hatten.

Der im ersten Lauf nicht zugeordnete Deskriptor „Weißrussland“ konnte dadurch ermittelt werden, dass ein automatischer Konverter von der alten zur neuen Rechtschreibung zugeschaltet war.

Die Ausgrenzung von Allgemeinwörtern sowie die allgemeine Begrenzung der Deskriptorenzahl wird durch eine vom Benutzer parametrisierbare Ausgabefunktion geregelt werden, die derzeit noch in der Entwicklung ist.

Hiermit kann dann das Stadium einer „tragbaren Verschmutzung“ der automatisch erzeugten Indexate mit unkorrekten/unerwünschten Elementen erreicht werden.

5 Ausblick: Forschung und weitere Experimente

Weitere Verbesserungen sind möglich durch Maßnahmen, die längerfristig angelegt sind, wo z.T. auch noch Forschung notwendig ist.

Eine komplexe Auswertung der im Thesaurus vorhandenen assoziativen Beziehungen („Related terms“) kann zu einer genaueren Gewichtungsberechnung für die Deskriptoren genutzt werden; hierzu gehört auch eine generelle Überarbeitung des Gewichtungsalgorithmus, der in Abhängigkeit von der Textlänge und auch von der Textart (Volltext oder Abstract) unterschiedlich gestaltet werden muss.

Bei einem routinemäßigen Einsatz wäre für die einzelnen Textsorten zu entscheiden, welche Teile der Dokumente in die Auswertung einzubeziehen wären. Bei grauer Literatur (Papers) bringen die langen Texte zuviel Vorschläge nicht-sintragenden Vokabulars. Bei Konzentration auf besonders inhaltsträchtige Teile des Textes sollte die Passgenauigkeit der Vorschläge zunehmen; als Vorschläge wurden z.B. die ersten und letzten 10% eines Textes diskutiert.

Aus den Handreichungen der „Sacherschließungsregeln“ wäre eine Modellierung der „Kettenbildung“ möglich, die das parallele Indexieren mit mehreren logischen Verknüpfungen erlaubt; hierzu müsste jedoch eine vollständige syntaktische Analyse der Texte durchgeführt werden. Es ist zu prüfen, inwieweit der notwendige (hohe) Aufwand die zu erzielenden Verbesserungen rechtfertigt.

Ähnliches gilt für die Bearbeitung von Texten, die einen hohen Anteil an fremdsprachlichen (meist englischen) Elementen aufweisen; hier wäre zu überlegen, ob solche (der deutschen Analyse unbekannt) Strings mit einer englischen Analyse bearbeitet werden sollen.

Als ein weiteres Experiment wäre die Bearbeitung einer größeren Menge von Dokumenten denkbar, die ganz in englischer Sprache vorliegen, was in der wirtschaftswissenschaftlichen Literatur häufig der Fall ist. Dies kann entweder auf der Basis des derzeit am IAI entwickelten bilingualen BINDEX-Verfahrens geschehen – oder aber auf der Basis eines ins Englische übertragenen STW-Thesaurus.

Ein (experimenteller) routinemäßiger Einsatz automatischer Indexierungsverfahren (deutsch und/oder englisch) könnte sich zukünftig anbieten z.B. zur Herstellung von Abstracts, zur Vor-Indexierung, die manuell nachgearbeitet würde, zur Indexierung elektronischer Papers, die aus Kapazitätsgründen sonst gar nicht verschlagwortet werden könnten; auch diese könnten ggf. manuell nachgearbeitet werden.

Insgesamt müsste der jetzt für die intellektuelle Indexierung eingerichtete STW kontinuierlich an die Erfordernisse einer automatischen Verschlagwortung angepasst werden. Das könnte auch hilfreich sein für seinen späteren potenziellen Einsatz im Rahmen einer domänenspezifischen Suchmaschine.

6 Referenzen

- [1] **Bakel van, Bas.** *Modern Classical Document Indexing.* In: Proceedings of SIGIR '98.
- [2] **Dumois, Susan.** *Latent Semantic Indexing: TREC-3 Report.* Proceeding of the Third Text Retrieval Conference, 1994.
- [3] **Landauer, Thomas K., Michael L. Littman.** *A statistical method for language-independent representation of the topical content of text segments.* Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research, 1990.
- [4] **Maas, Heinz Dieter.** *Multilinguale Textverarbeitung mit Mpro.* In: G. Lobin et al. (eds.): Europäische Kommunikationskybernetik heute und morgen. KoPäd, München, 1998.
- [5] **Maas, Heinz Dieter.** *Thesaurus als Wissensbasis für Begriffszerlegungen.* Information. Proceedings, Friedrich-Schiller-Universität Jena, 28. bis 30. September 1993.
- [6] **Plaunt, Christian, Barbara A. Norgard.** *An Association Based Method for Automatic Indexing with a Controlled Vocabulary,* Technical Paper, University of California at Berkeley
- [7] **Salton, Gerald.** *Automatic Text Processing.* Addison-Wesley Publishing, 1989.
- [8] **Preissuchen – WWW –Suchmaschinen, Kataloge und Metasucher im Vergleich,** in: c't 23/99, pages 162ff.
- [9] **Haller, Johann.** *Die Erschließung natürlichsprachlicher Information im Informationssystem CONDOR,* in : Nachrichten für Dokumentation 4/5-1978