In: Knorz, Gerhard; Kuhlen, Rainer (Hg.): Informationskompetenz – Basiskompetenz in der Informationsgesellschaft. Proceedings des 7. Internationalen Symposiums für Informationswissenschaft (ISI 2000), Darmstadt, 8. – 10. November 2000. Konstanz: UVK Verlagsgesellschaft mbH, 2000. S. 71 – 87

Wortmodell und Begriffssprache als Basis des semantischen Retrievals

Gerhard Rahmstorf
Oberer Rainweg 57
69118 Heidelberg
Tel. 06221-808129
Fax. 06221-802682
rahmstorf@regio-info.de
Technische Universität Darmstadt

Zusammenfassung

Der heutigen Retrievaltechnik wird das Projekt eines semantisch basierten Suchsystems gegenübergestellt. Es soll genauer und vollständiger arbeiten sowie systematische Zusammenhänge zwischen Themen unterstützen. Bei diesem Ansatz wird ein umfassendes Wörterbuch mit einer einfachen begrifflichen Darstellung der Wortbedeutungen benötigt. Das Wortmodell bildet Wort, Wortmerkmale, Lemma, Wortbedeutungen (Lesarten), Lesartenmerkmale und Begriffe ab. Begriffe sind formale Ausdrücke einer Begriffssprache. Entsprechend dieser Differenzierung wird Lemmaindexierung, Lesartenindexierung und Begriffsindexierung unterschieden. Begriffe werden mit dem Programm Concepto grafisch konstruiert und erfasst.

Abstract

The requirements for a more perfect retrieval are described from the user's viewpoint: thematic search, queries in the form of natural language phrases, basic interpretation of word meanings etc. A word model for such a system supports three different types of objects: words, readings and concepts. Concepts are expressions of a concept language. A software system called Concepto supports the acquisition of lexical data and the graphical construction of formal concept expressions.

1. Einleitung

Schon früh wurde in der Geschichte der Informationstechnik vom semantischen Retrieval gesprochen[1]. In der Praxis hat sich jedoch der schon in den frühen sechziger Jahren etablierte Ansatz des zeichenkettenbasierten Recherchierens durchgesetzt. Dieser Ansatz dominiert in verschiedenen Varianten die heutige Suchtechnik. Die Varianten unterscheiden sich z. B. bezüglich der Funktionen bei der Abfrage (Boolesche Operatoren), der für die automatische oder intellektuelle Dokumentcharakterisierung verwendeten Basis (Trigramme, "Wörter", Deskriptoren eines kontrollierten Vokabulars), der textuellen Vorverarbeitungen (Erkennung von Synonymen und Wortformen) und weiterer nachgeschalteter Funktionen (Relevanzfeedback). Das Spektrum der Methoden wird verfeinert und auch für die Besonderheiten der Suche im Web weiterentwickelt (Mladenic 1999). Diese Verfahren können in differenzierter Weise Relevanz von Dokumenten bestimmen, allerdings auf der Basis des Vorkommens von Zeichenketten und der formbezogenen Eigenschaften der Anfrage und der Dokumenttexte. Erst wenn ein Thesaurus bei der Relevanzanalyse verwendet wird, können wenige semantische Beziehungen bei der Relevanzanalyse berücksichtigt werden.

Die eingeführte Retrievaltechnik hat den großen Vorteil, dass mit ihr schnell und kostengünstig große Mengen von Texten automatisch indexiert werden. Sie ist robust, gerade weil sie ohne linguistische Analysen und komplexe Begriffsstrukturen auskommt. Das Schlagwort "Volltextrecherche" signalisiert, dass ohne nennenswerte intellektuelle Zusatzarbeit jeder beliebige Text beliebiger

---

[1] Im Arbeitsgebiet Information und Dokumentation war von Beginn an klar, dass es beim Dokumentieren, Ordnen und Recherchieren primär um Inhalte und Themen geht. Daher wurden im Verlaufe der letzten 30 Jahre immer wieder Retrievalsysteme entworfen und erprobt, bei denen natürliche Sprache, semantische Analysen, begriffliche Repräsentationen, Weltwissen und Inferenztechniken eine Rolle spielen. Eine angemessene Würdigung dieser Entwicklungen würde den Rahmen dieser Arbeit sprengen. Das semantische Retrieval in der hier dargestellten Form ist nur eine Variante unter den Entwicklungen, die das klassische Schema verlassen. Es geht auf Überlegungen des Verfassers aus den frühen siebziger Jahren zurück. Zu dieser Zeit wurden schon roles und links als Instrumente zur Entwicklung einer genaueren Indexierung und Recherche diskutiert. Die Linguistik hatte mit der Kasusgrammatik (Fillmore 1968) einen inspirierenden Beitrag zur semantischen Strukturierung geliefert. Tiefenkasus ließen sich in semantischen Netzen verwenden. Die Idee des Indexierens und Recherchierens mit Phrasen und der Gedanke der Relevanzanalyse auf der Grundlage der begrifflichen Phrasenrepräsentationen wurden in (Rahmstorf 1978) skizziert. Die Bezeichnung "Semantic Information Retrieval" hat schon Raphael 1968 für sein System verwendet, das jedoch nicht die hier dargestellte Methodik verfolgt. Schank bezeichnete mit "Conceptual Information Retrieval" eine sehr viel weitergehende Aufgabenstellung, zu der die Analyse von Texten, die konzeptuelle Repräsentation von Fakten aus diesen Texten, die Pflege der Daten und die Beantwortung von frei formulierten Anfragen gehören (Schank, Kolodner, DeJong 1980). Die unbeschränkte Verwendbarkeit der natürlichen Sprache stellt bis heute eine schwer zu nehmende Hürde dar. Dokumentation und Retrieval haben sich weitgehend auf Wörter (Deskriptoren) beschränkt. Der vielfältige Einsatz von linguistischen Mitteln (Klavans 1994) und die Versuche, Methoden der Künstlichen Intelligenz und Expertensystemtechnik für bestimmte Aufgaben des Retrievals anzuwenden, führten bisher nicht zu einem neuen praktikablen Paradigma des Retrievals.

Sprachen maschinell recherchierbar gemacht werden kann. Das ist daher -soweit betrachtet- die passende Methode für den immensen Zuwachs an maschinell verfügbarem Material im Internet. Alternative Methoden haben nur deshalb eine Chance, in Zukunft verwendet zu werden, weil die Genauigkeit und Vollständigkeit der Recherchen unbefriedigend sind, und zwar insbesondere dann, wenn nach spezifischen Themen gesucht wird, die nicht durch Eigennamen oder andere eindeutige Codes gesucht werden können.

Ein weiterer Einwand gegen die heutige Recherchetechnik ist die fehlende Ordnung und Durchschaubarkeit des Verfahrens und der Ergebnispräsentation. Zumindest das Fachpublikum möchte Zusammenhänge, Kontexte und Ordnung in der Masse der angebotenen Dokumente erkennen. So wie man von der Buchwelt herkommend wohlgeordnete Lehrwerke, Fachwörterbücher, Handbücher, Enzyklopädien und systematische Vertiefungen in Form von Monographien kennt, so möchte man auch in der elektronischen Welt bessere Hilfen zur Orientierung haben und ein systematisch gestaltetes Wissenssystem benutzen können.

Alternative Retrievalverfahren unterscheiden sich von dem etablierten Ansatz nicht nur durch die Einführung von Begriffen als Basis für die Bestimmung der Relevanz eines Dokuments bezogen auf einen Text, sondern auch durch die Einführung einer Informationssprache für Recherche und Indexierung. Diese Sprache umfasst Wörter, feste Phrasen und eine offene Menge regelhaft gebildeter Benennungen, vorwiegend vom Typ Nominalphrase (Tabelle 1).

Informationssprache	**ohne Semantik: Zeichenketten**	**mit Semantik: Begriffsdarstellung**
sublexikalische Einheiten	n-gramme	-
lexikalische Einheiten: kontrolliert	Thesaurus: Deskriptoren	(Notationen, Thesaurus online für Recherche)
lexikalische Einheiten: beliebige Wörter	Gegenwärtige Technik: "Freitextsuche"	
syntaktische Einheiten: kontroll. Phrasenmenge	(Thesaurus) begrenzt möglich	
syntaktische Einheiten: freie Phrasensprache		semantisches Retrieval

Tabelle 1: Retrievaltechniken

Ordnung wird bei diesem Ansatz dadurch erzielt, dass die Begriffsdarstellung soweit möglich und zweckmäßig dem folgt, was die Benennungen besagen. Die Relevanz wird auf der Basis von Beziehungen zwischen Begriffen bestimmt. Der Entwicklungsaufwand muss sich lohnen. Wird ein qualitativ besseres Retrieval wirklich dringend benötigt? Die Benutzer können das heute ebensowenig beantworten, wie sie vor 15 Jahren die Frage hätten beantworten können, ob sie ein Internet benötigen würden? Wir müssen uns daher selber in die Rolle eines Benutzers versetzen und den Ansatz des semantischen Retrievals aus den

Interessen der informationssuchenden Anwender, insbesondere der Fachleute aus Wissenschaft, Technik und Wirtschaft, begründen.

Die Anwender haben sich ebenso wie die Informationsexperten an die Arbeitsweise von Datenbanken und Suchmaschinen gewöhnt. Sie kennen nichts anderes. Sie wissen auch oft, wie die Programme ihre Suchresultate erzielen. Sie kommen daher gar nicht erst auf den Gedanken, von einem System die perfekte Bereitstellung von gesuchtem Wissen zu fordern und an dieses System wie an einen fachkundigen Bibliothekar heranzutreten. Bei Gesprächen mit Bibliothekaren werden sie nicht verzichten wollen auf thematische Suche, freie Wahl der Themenbenennung, Unabhängigkeit von festgesetzten Strukturierungen und veralteten Systematiken u. a.

2. Anforderungen 2.1 Thematische Suche

Benutzer suchen nach Literatur zu einem Thema. Diese thematische Suche unterscheidet sich von anderen Arten der Suche, z. B. der spontanen Entscheidung für angebotene Links in einem Hypertext, der Suche nach einem bestimmten, schon bekannten Dokument oder der Suche nach Dokumenten, in denen bestimmte Wörter oder Zeichenketten vorkommen. Bei der thematischen Suche sind die Dokumente für den Benutzer relevant, die das von ihm angegebene Thema behandeln, und zwar unabhängig von inhaltlichen Einzelheiten, Stil, Gestaltung, Gliederung und Wortwahl der möglichen Texte.

2.2 Unabhängigkeit von Strukturfestlegungen

Benutzer möchten nicht unbedingt in einer bestimmten Struktur suchen, die in Form eines Menübaums oder eines bibliothekarischen Klassifikationssystems vorgefertigt wurde. Ordnungssysteme können dem Benutzer Orientierung vermitteln oder auch zur Navigation dienen. Sie sollten aber möglichst die bedeutungsgemäßen Beziehungen zwischen den Klassenbenennungen wiedergeben.

2.3 Natürliche Sprache für die Anfrage

Benutzer möchten das Thema mit den differenzierenden Mitteln der eigenen Sprache ausdrücken können. Sie wollen die Wörter verwenden, die ihrer Meinung nach der Sache angemessen sind. Sie wollen sich nicht auf bestimmte Wörter oder Codes beschränken, die die Suchlogik des Systems unterstützt. Thesauren und andere kontrollierte Vokabulare stellen eine Erschwernis dar, die insbesondere dem Internetbenutzer nicht ohne Not zugemutet werden sollte.

Zur Beschreibung eines Themas sollte man die sprachlichen Konstruktionen verwenden können, die das jeweilige Thema so präzis wie nötig gegen andere Themen abgrenzen. Zur Themenbenennung sollte die Retrievaltechnik daher diejenigen Attributformen unterstützen, die zur Benennung für Titel in der fachlichen Literatur verwendet werden. Das sind z. B. Adjektivattribute, Genetivattribute und Präpositionalattribute, die ein substantivisches Bezugswort

näher bestimmen. Erst wenn solche Phrasen als Ausdrücke der Informationssprache zugelassen werden, wird klar, welches Thema den Benutzer eigentlich interessiert. Die Relevanz von Dokumenten bezüglich einer Anfrage kann nur dann genauer bewertet werden, wenn die Anfrage explizit ausformuliert ist und sich damit von anderen Anfragen, die mit den gleichen Wörtern gebildet wurden, unterscheidet. Durch die sprachliche Formulierung werden die Wörter nicht nur in eine bestimmte Reihenfolge gebracht, sondern auch syntaktisch strukturiert. Mit Hilfe dieser Struktur wird der Ausdruck interpretiert. Aus den drei Wörtern *Ausbildung, Arbeitsloser* und *Computertechnik* können z. B. die folgenden Phrasen unterschiedlicher Bedeutung gebildet werden (Rahmstorf 1994, Lein 1994):

- Ausbildung der Arbeitslosen in der Computertechnik
- Ausbildung in der Computertechnik durch Arbeitslose
- Computertechnik zur Ausbildung für Arbeitslose
- Arbeitslose mit Ausbildung in Computertechnik
- Arbeitslose in der Ausbildung für Computertechnik

Phrasen stellen nur einen Teil der Ausdrucksmöglichkeiten der natürlichen Sprache dar. Zu dieser Informationssprache gehören z. B. keine vollständigen Sätze.

3. Komponenten des semantischen Retrievals

Mit den natürlichen Benutzeranforderungen wird der Rahmen für das semantische Retrieval festgelegt. Diese Retrievaltechnik wird kurz durch drei Eigenschaften gekennzeichnet: sie bietet die Möglichkeit, mit Phrasen zu indexieren und zu recherchieren, sie stellt die Phrasen als begriffliche Strukturen dar und sie bestimmt die Relevanz von Texten für eine gegebene Anfrage auf der Basis der begrifflichen Repräsentation des Anfragethemas und der Themen der Informationsangebote. Dabei werden gebietsunabhängige Relevanzregeln verwendet.

Abbildung 1 stellt den symmetrischen Aufbau für den Datenfluß im semantischen Retrieval dar. Die anzubietenden Dokumente oder Texte werden in eine Dokumentdatenbank geladen (Prozeß DG). Den Texten oder Kapiteln eines Textes müssen Themen durch jeweils eine Phrase zugewiesen werden. Diese Indexierung muß ein Indexierer machen (HI), solange noch keine Programme vorhanden sind, die ein automatisches Phrasenindexieren (AI) ermöglichen.

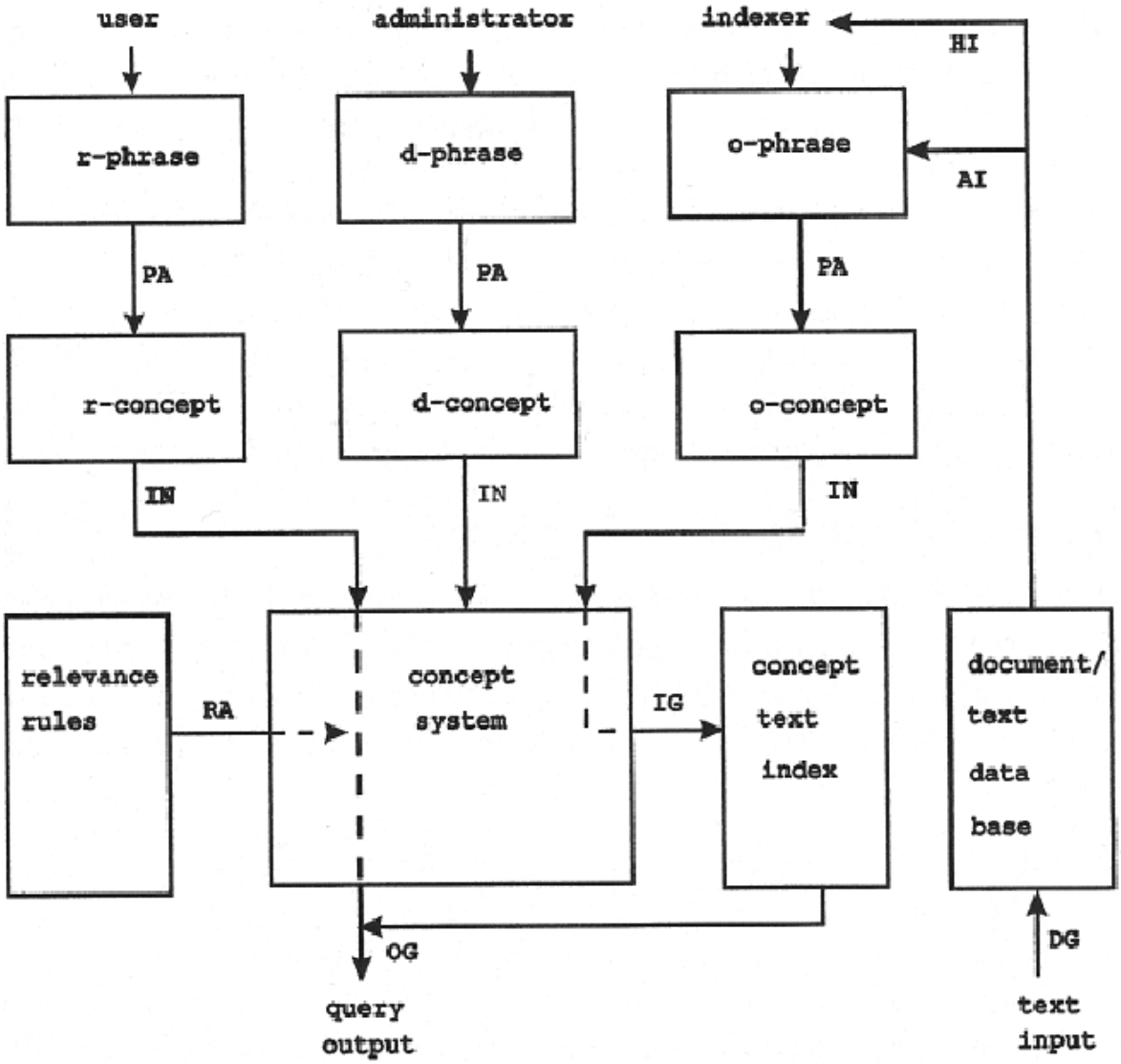

Abbildung 1: Systemübersicht zum semantischen Retrieval

Der als offering phrase (o-Phrase) bezeichnete Ausdruck gehört einer der vom System unterstützten natürlichen Sprachen an. Er muss lexikalisch, syntaktisch und semantisch analysiert werden (PA), um als Begriff dargestellt werden zu können. Der Begriff wird als Struktur in das Netz der schon vorhandenen Begriffe eingefügt (IN). Zu diesem Begriff wird ein Verweis bzw. Index auf den zugehörigen Text generiert (IG). Die Verweise bilden zusammen einen Index für eine gegebene Textdatenbank. Bei der Recherche kann der Benutzer seinen Informationsbedarf mit einer Phrase ausdrücken. Diese request phrase (r-Phrase) wird danach wie die o-Phrase analysiert (PA). Die Anfrage wird als Begriff im vorhandenen Begriffsnetz integriert (IN). Die Relevanzanalyse (RA) bestimmt, welche o-Begriffe für die gegebene Anfrage relevant sind. Für die relevanten Begriffe werden die zugehörigen Texte zusammengestellt (OG) und dem Benutzer ausgegeben.

Die begriffliche Basisstruktur, die mit dem Wortschatz einer Sprache verbunden ist, muss in dem Retrievalsystem intellektuell erfasst werden. Ausgangspunkt jeder Bedeutungsbeschreibung bilden einfache verbale Definitionen der einzelnen Wortbedeutungen (Lesarten). Auch die Definitionen sind syntaktisch betrachtet Phrasen. Durch diese d-Phrasen (definition phrase) werden Begriffe auf andere Begriffe zurückgeführt. Alle Begriffe bilden ein Netz, das zur Relevanzanalyse verwendet wird. Die Beziehungen zwischen Begriffen legen Zugriffswege fest. Wenn z. B. *Sehen* als *visuelles Wahrnehmen* definiert wird, dann können Texte, die sich mit dem Thema *Sehen* befassen, auch über die Begriffe *Wahrnehmung* und *visuell* erreicht werden. Die Basisstruktur wird zusammen mit dem Wörterbuch betreut und aktualisiert. Diese Aufgabe ist mit der Pflege eines Thesaurus vergleichbar. Die Benutzer müssen sich mit dieser Struktur nicht befassen.

3.1 Anforderungen an das Wörterbuch

Wörterbücher für semantisches Retrieval müssen u. a. folgende Aspekte der Benennungen unterstützen: Lesarten (Homonymität), Synonymität, Wortäquivalenz (Übersetzung) und begriffliche Strukturen. Dazu muss ein entsprechend differenziertes Wortmodell für die lexikalischen Daten aufgebaut werden.

Wenn die Benutzer ihre Themen frei benennen können, aber ein beträchtlicher Teil der Wörter natürlicher Sprachen mehrdeutig ist, müssen die verschiedenen Lesarten der Wörter im Wörterbuch des Retrievalsystems unterschieden werden. Es gehört aber auch zu den Eigenarten aller natürlichen Sprachen, dass ein bestimmtes Thema mit verschiedenen Wörtern und Ausdrücken benannt werden kann. Die Benutzer wollen sich aber nicht auf bestimmte Vorzugsbenennungen, die die Maschine erkennt, festlegen lassen. Daher müssen verschiedene Ausdrücke, die die gleiche Bedeutung haben, durch eine einzige begriffliche Repräsentation im Retrievalsystem dargestellt werden.

Auch in einer globalisierten Welt muss die Telekommunikation und das Recherchieren auf der Basis der verschiedenen Einzelsprachen durchführbar sein. Wenn mehrsprachiger Zugriff möglich sein soll, muss das Retrievalsystem über Wörterbücher für jede unterstützte natürliche Sprache verfügen. Äquivalente Wörter und Benennungsausdrücke von verschiedenen Sprachen müssen als gleichbedeutend erkannt und dargestellt werden.

Jeder Benutzer weiß, dass zwischen den unendlich vielen Themen, zu denen Information angeboten und gesucht werden kann, Beziehungen bestehen. In der Kommunikation zwischen Benutzer und Bibliothekar sind diese Beziehungen Grundlage des gegenseitigen Verstehens. Im Retrievalsystem sollten sie annähernd abgebildet werden. Bedeutungsstrukturen sollen nicht unnötig komplex werden, damit die begriffliche Repräsentation in technische Hinsicht und bezüglich der Erstellung von Strukturen beherrschbar bleibt. Das Wortmodell wird daher durch eine verhältnismäßig einfache Begriffssprache ergänzt. So wie in der natürlichen Sprache die Grammatik den Rahmen für die Satzbildungen festlegt,

legt die Begriffssprache den Rahmen für die begrifflichen Strukturen fest, die den einzelnen Lesarten eines Wortes zugewiesen werden können.

3.2 Wortmodell

Aus den genannten Anforderungen leitet sich eine differenzierte Modellierung der sprachlichen Gegebenheiten ab. In dem Retrievalsystem müssen vier Arten von linguistischen Objekten unterschieden werden: Sprachen, Wörter, Wortbedeutungen (Lesarten) und Begriffe. Diese Objekte haben jeweils eigene Eigenschaften (Attribute). Zwischen den Objekten bestehen bestimmte Arten von Beziehungen. Sprachen sind hauptsächlich durch ihren Wortschatz und ihre Grammatik bestimmt. Für die Modellierung ist zu beachten, dass sich die unterschiedlichen Grammatiken der Einzelsprachen auch darin niederschlagen, dass die Kategorien der Wortbeschreibung in den Einzelsprachen voneinander abweichen. Wir können daher nicht von einem einheitlichen Wortbeschreibungsschema für alle Sprachen ausgehen. Die den Wörtern zuzuweisenden Attribute, z. B. Genus, Deklinationsklasse, und die jeweils möglichen Attributwerte sind sprachspezifisch. Mehrsprachig betreibbare Retrievalsysteme, die auch für Übersetzungsfunktionen verwendet werden, müssen daher eine an die jeweilige Sprache anpassbare Merkmalstruktur für die Wortbeschreibung aufweisen.

Zu den Einheiten des Wortschatzes gehören nicht nur die "Wörter", sondern auch Mehrwortausdrücke, die eine spezifische, nicht regelhaft rekonstruierbare Bedeutung haben. Jedes Wort gehört zu nur einer Einzelsprache. Wörter können in verschiedenen Formen vorkommen. Die Wortidentität wird durch die Benennungsform des Wortes und durch weitere Merkmale bestimmt. Die Zeichenkette *Kiefer* als Lemma reicht nicht aus, um das Wort eindeutig zu identifizieren. In diesem Fall ist der Wert des Merkmals Genus zusätzlich anzugeben: *die Kiefer* bezeichnet etwas anderes als *der Kiefer.* Im Englischen *muss* die Zeichenkette *capital* durch einen Wert für die Kategorie Wortart - Nomen oder Adjektiv - identifiziert werden. Das Lemma als die Zeichenkette, mit der ein Wort benannt wird, ist im Deutschen nach der Orthographiereform durch die inflationäre Zulassung von Schreibvarianten kein eindeutiges Wortmerkmal mehr. Eine datenbankmethodisch saubere Abbildung dieser weiteren 1:n-Beziehung wird unzumutbar komplex. Man wird daher in einem Computerwörterbuch für Retrieval und Übersetzung entweder benutzungserschwerende Verweise oder redundante Einträge für die Schreibweisen *Fotografie, Photographie* und eventuell weitere Varianten haben. Dies ist deshalb ein Problem, weil das Lemma für die Praxis verschiedene Funktionen erfüllt. Es ist Aufrufbenennung eines Wörterbucheintrags, Grundlage der alphabetischen Einsortierung der Wörter, gemeinsame Kennzeichnung aller Lesarten eines Wortes in der Strukturvisualisierung u. a.

Wörter können mehrere Lesarten haben. So hat z. B. das englische Substantiv *capital* vier Lesarten. Jede Lesart wird durch Lesartenmerkmale beschrieben. Die in der Praxis wichtigsten Lesartenmerkmale sind: verbale Definition, Markierungen bezüglich Sprachebene, Metaphorik u. a., Beispielsatz zur Charakterisierung von

Bedeutung und Gebrauchsweise, Fachgebietsangabe, kommentierende Hinweise, Lesartenidentifikation. Die Kategorien, mit denen Lesarten beschrieben werden, sind für alle Sprachen gleich. Zumindest sind derzeit keine Gründe erkennbar, warum für die informationstechnische Anwendung sprachspezifische Unterschiede in der Lesartenkategorisierung vorgesehen werden sollten. Die Lesartenmerkmale werden im Datenmodell sauber von den Wortmerkmalen unterschieden. Diese Gegebenheiten der Sprache machen die Benutzungsoberfläche bei der Wortschatzerfassung komplexer. Die Vorteile dieser lexikographischen Oberfläche möchten aber die Entwickler des Wörterbuchs nach einer kurzen Einarbeitungszeit nicht mehr missen.

Ein wichtiger Modellierungsaspekt ist, dass jede Lesart ihre eigene Identität hat. Dies soll am Beispiel der Lesarten der Wörter *Anfang* und *Beginn* gezeigt werden. Jedes dieser beiden Wörter hat nur eine Lesart. Die Wörter sind synonym, d. h. ihre Bedeutungen stimmen überein. Dennoch sind die Beschreibungen der beiden Lesarten unterschiedlich. Der Beispielsatz muss unterschiedlich sein. Die verbale Definition kann unterschiedlich formuliert sein. Die wichtigsten Daten zur Charakterisierung der Lesarten sind textueller Art und damit sprachspezifisch. Deutsche Wörter werden normalerweise durch deutsche Definitionen beschrieben, französische Wörter durch französische Bedeutungsumschreibungen. Daher können zwei Lesarten, die Wörtern verschiedener Sprachen angehören, nicht identisch sein, selbst wenn sie ihrer Bedeutung nach äquivalent sind.

Jeder Lesart kann mindestens ein Begriff zugewiesen werden. Die Lesart ist ein sprachspezifisches Objekt, der Begriff ist sprachinvariant. Lesarten können nicht ohne Wörter, aber ohne Begriffe spezifiziert werden. Begriffe können durch formale Ausdrücke spezifiziert werden, ohne einer Lesart eines Wortes zugewiesen sein zu müssen. Ein gegebener Begriff kann verschiedenen Lesarten zugewiesen sein. Dies wird immer dann der Fall sein, wenn diese Lesarten von Wörtern einer bestimmten Sprache als synonym gelten. Entsprechendes gilt für Lesarten, die Wörtern unterschiedlicher Sprachen angehören. Man spricht dann von einer interlingualen Synonymität oder auch von semantischer Äquivalenz. Die Begriffe werden durch die noch zu beschreibende Begriffssprache charakterisiert. Zwischen Wörtern und Begriffen bestehen semantische Beziehungen. Zwischen Begriffen bestehen begriffliche Beziehungen.

Das hier beschriebene Wortmodell wurde in dem Programm Concepto implementiert. Dieses Programm ist kommerziell verfügbar von Antje Rahmstorf Sprachsysteme, 69118 Heidelberg, Oberer Rainweg 57. Es unterstützt auch die Begriffssprache und die dazugehörigen Operationen der grafischen Konstruktion von Begriffen. In dem Wortmodell werden folgende Objekte, Relationen und Attribute unterschieden:

1. Sprache

2. Benennungseinheit (Wort, idiomatische Phrase, frei formulierte Benennung)

3. Merkmale des Wortes für sprachspezifische Kategorien

4. Lemma als besonderes Merkmal des Wortes
5. Lesarten
6. Merkmale der Lesarten
7. Begriffe, die Lesarten zugewiesen werden
8. Beziehungen zwischen Begriffen

Die Objekte müssen aus technischen Gründen identifizierbar sein. Es fällt wegen der Wortmerkmale schwer, für Wörter eine bestimmte Identifizierung, z. B. durch eine vereinbarte Wortnummer, durchzuführen. Jede Institution hat daher seine eigenen Wortnummern. Lesarten lassen sich dagegen leichter identifizieren. Eine Lesartennummer ist notwendig, um eine klare Schnittstelle zwischen Lesart und Begriff einzuführen. Die Begriffe identifizieren sich durch die formalen Ausdrücke der Begriffssprache.

3.3 Begriffssprache

Mit einer Begriffssprache lassen sich Bedeutungen von Benennungsausdrücken der Einzelsprachen in einer sprachübergreifenden formalen Weise darstellen. Jede Begriffssprache beschreibt ein Spektrum von möglichen Strukturen oder Repräsentationen. Bei dem Aufbau von begrifflichen Strukturen geht man von den verbalen Definitionen der einzelnen Lesarten eines Wortes aus. Nicht jede Lesart eines Wortes kann definiert werden. Undefinierbare Wörter bzw. Lesarten werden in der Begriffssprache als undefinierte Begriffe eingeführt. Sie erhalten keine definierende Struktur. Um die undefinierten Begriffe, die für den Menschen verschiedene Bedeutungen haben, in der Maschine unterscheiden zu können, kann man ihnen jeweils andere Notationen der Begriffssprache als Formalausdrücke zuweisen. Nicht jede Lesart, die verbal definiert werden kann, muss auch in der Begriffsprache formal definiert werden. Definierte Strukturen werden nur dann erstellt, wenn sie für die jeweilige Anwendung Vorteile bringen.

Begriffliche Strukturen unterstützen Recherchen, Schlussfolgerungen, Übersetzungen u. a.. Sie werden von Informatikern, Informationsfachleuten und anderen Experten benutzt. Benutzer dieser Systeme müssen die begrifflichen Strukturen nicht unbedingt verstehen. Sie können sie aber zur Navigation verwenden. Dennoch dürfen Begriffssysteme nicht in eine undurchschaubare, unkontrollierbare Komplexität ausufern. Sie müssen für den Entwickler der Struktur leicht interpretierbar sein. Hinzu kommt, dass formale Ausdrücke einer Begriffssprache in textueller Form schwerer verständlich sind. Eine Visualisierung der formalen Ausdrücke ist für die Erfassung der begrifflichen Strukturen notwendig. Das erwähnte Programm Concepto unterstützt alle notwendigen Funktionen der Editierung und grafischen Manipulation von begrifflichen Strukturen.

Unter dem Anwendungsaspekt der Recherchetechnik betrachtet müssen Begriffssprachen für beliebige Benennungen aus allen denkbaren Gebieten offen sein und nicht von den Besonderheiten der jeweiligen Fachgebiete geprägt werden. Die Grenzen zwischen den Fachgebieten sind ohnehin fließend. Wer in der Biochemie tätig ist, will mit denselben Recherchetechniken in der Chemie, der Biologie, der Medizin und der Informatik suchen können. Ziel einer begrifflichen Repräsentation in der Informationstechnik ist es daher, sprachinvariant und fachgebietsinvariant zu sein. Dies wird u. a. dadurch erreicht, dass die Relationstypen, die in der Begriffssprache CLF von Concepto verwendet werden, nicht aus Fachgebieten, sondern aus den Attributformen, Wortbildungsmustern und anderen universalen grammatischen Mitteln abgeleitet worden sind.

3.3.1 Begriffliche Repräsentation

Abb. 4 zeigt als Beispiel eine begriffliche Repräsentation für die Wörter *Energie* und *Transport* und für weitere Begriffe im Umfeld dieser Wörter. Die Knoten stellen Begriffe dar, die Kanten symbolisieren begriffliche Beziehungen. Die Wörter *Energie* und *Transport* sind durch folgende Phrasen definiert worden:

Energie := Fähigkeit, Arbeit zu leisten

Transport := Arbeit, die den Zweck hat, den Standort von Objekten zu verändern

Die Analyse dieser Definitionsphrasen führt zu der begrifflichen Struktur, die im Diagramm oberhalb der Begriffe *Energie* und *Transport* dargestellt ist.

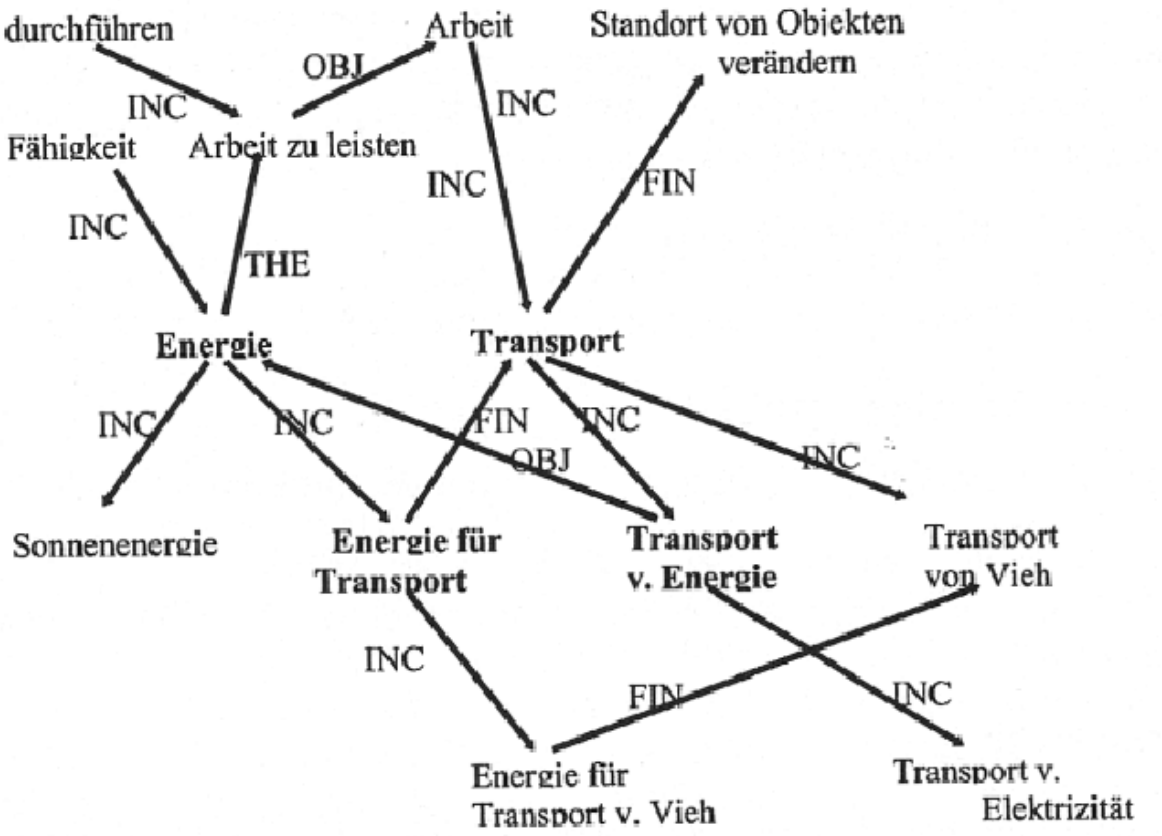

Abbildung 3: Beispiel einer begrifflichen Darstellung

In dieser Begriffsrepräsentation werden wie in einem Thesaurus nur Beziehungen aus einem begrenzten Vorrat an zweistelligen Relationen zugelassen. Die im Beispiel vorkommenden Beziehungen und ihre Paraphrasen sind:

Inklusionsrelation	INC(x,y)	y ist ein x
Finalrelation	FIN(x,y)	y ist der Zweck von x
Objektrelation	OBJ(x,y)	y ist das Objekt von x
Themarelation	THE(x,y)	y ist das Thema von x

Jeder Begriff wird hier durch einen direkten Oberbegriff und durch einen direkten Differenzbegriff als begriffliche Struktur definiert. Der Begriff *Fähigkeit* definiert mit der Inklusionsrelation INC den Begriff *Energie. Fähigkeit* ist daher der direkte Oberbegriff von *Energie. Arbeit leisten* ist der direkte Differenzbegriff von *Energie.* Zum Umfeld eines definierten Allgemeinbegriffs x gehören außer diesen zwei definierenden Positionen direkter Oberbegriff o(x) und direkter Differenzbegriff d(x) die folgenden fakultativen Positionen: direkter Unterbegriff u(x) und direkter Seitenbegriff s(x). Wenn y=o(x), dann ist x=u(y). Ebenso gilt: wenn y$^{=}$d(x), dann ist x=s(y). So ist z. B. der Begriff *Energie* ein Seitenbegriff von *Arbeit leisten,* weil *Energie* mit dem Begriff *Arbeit leisten* als Differenzbegriff definiert wurde.

Die beiden Phrasen *Transport von Energie* und *Energie für Transport* können entsprechend ihrer unterschiedlichen Bedeutung durch zwei verschiedene Begriffe explizit dargestellt werden:

Energie für Transport = u(Energie)

Energie für Transport = s(Transport) mit der Differenzrelation FIN

Transport von Energie = u(Transport)

Transport von Energie = s(Energie) mit der Differenzrelation OBJ

Im nichtsemantischen Retrieval auf Wortbasis würden jede der beiden Phrasen auf die Menge der beiden Deskriptoren {Energie, Transport}, zurückgeführt.

3.3.2 Indexierung

Der symmetrische Ansatz des semantischen Retrievals fordert eine gleichartige Vorgehensweise für Anfrage und Indexierung. Wenn thematisch gesucht werden soll, wird auch eine thematische Indexierung benötigt. Man kann prinzipiell auf den Ebenen der Benennung, der Lesarten und der Begriffe indexieren. In Tab. 2 sind die diese drei Indexierungsmittel für zwei Indexierungsarten angegeben. Bei der ersten Art (Fälle A, B und C) beschränkt sich die Indexierung auf Mengen von Vokabulareinheiten. Das können Wörter oder Mehrwortausdrücke aus einer festgelegten Menge von Phrasen sein. Jede Einheit muss in einem gespeicherten Lexikon vorhanden sein. Die zweite Art (Fälle D, E und F) ist die interessantere. Hier steht für die Indexierung zusätzlich eine Benennungssprache zur Verfügung, die eine syntaktische Verknüpfung der Lexikoneinheiten zulässt.

	Indexate mit Mitteln der		
Indexierungssprache	**Ausdrucks-formen**	**Lesarten-beschreibung**	**Begriffs-darstellung**
nur Vokabular (Wörter und lexi-kalisierte Phrasen)	A Lemmata	B Lesarten-Nummer	C formale Ausdrücke
zusätzl. regelhaft gebildete Benennungen (Phrasen)	D. Phrase als Zeichenkette	E -	F formale Ausdrücke

Tabelle 2: Möglichkeiten der Indexierung

Der Fall A entspricht der vokabularbezogenen Indexierung mit den Lemmas, wie sie aus der gegenwärtigen Technik bekannt ist. Der Fall B berücksichtigt die Mehrdeutigkeit der Wörter bei der Indexierung. Da gegenwärtig keine

Lesartennummern in der Praxis eingeführt sind, werden bei kontrollierten Vokabularen Vorzugsbenennungen oder Lemmata mit Zusatzsymbolen verwendet, die für die jeweilen Lesarten zu vereinbaren sind. Statt dieser Mittel kann man für eine Dokumentbeschreibung durch Lesartenmengen auch die Lesartennummern verwenden, sobald eine Verständigung über ein solches System der Lesartenidentifikation erfolgt ist. Der Fall C stellt die Indexierung mit formalen Ausdrücken für Begriffe dar. Er hat gegenüber dem Fall B den Vorteil, dass die Indexierung nicht in einer bestimmten Einzelsprache (Deutsch, Englisch usw.) erfolgt.

Der Fall D besagt, dass einem Dokument als Indexat ein beliebiges Thema in Form einer sprachlichen Benennung als Zeichenkette zugewiesen wird. Damit kann zwar wesentlich präziser als mit Wortmengen indexiert werden, aber die Wiederauffindbarkeit solcher Dokumente kann nur durch eine entsprechende linguistische Analyse des Benennungsausdrucks erzielt werden. Mit der Zuweisung einer das Thema treffenden Phrase ist die informationswissenschaftlich kritische Indexierungsleistung erbracht, aber die informationstechnische Recherchierbarkeit noch nicht gelöst. Durch den nachfolgenden Analyseprozess muss der Zeichenkette die begriffliche Repräsentation als formaler Ausdruck entsprechend dem Fall F zugeordnet werden. Dafür stehen heute noch keine praxisreifen Lösungen zur Verfügung. Dessenungeachtet ist die Indexierung beliebiger Themen durch formale Ausdrücke der Begriffssprache das Endziel des semantischen Retrievals. Der Fall E stellt keine brauchbare Alternative dar, weil nicht alle Ausdrücke einer Indexierungssprache, die das große Vokabular der natürlichen Sprachen verwenden, durch Lesartennummern identifiziert werden können.

Mit dem Wortmodell und der Begriffssprache sind notwendige Voraussetzungen für eine neue Recherchetechnik beschrieben worden. Diese können mit dem Programm Concepto verwendet werden. Es geht jetzt darum, weitere Erfahrungen mit der Strukturierung von großen Mengen von Wörtern zu sammeln. Die von Concepto zugelassenen Möglichkeiten der Begriffssprache werden für den Aufbau eines umfassenden Begriffssystems für Retrieval in einer methodisch kontrollierten Weise eingesetzt. Dazu gehören Beschränkungen beim Aufbau von Begriffsstrukturen, Festlegung auf ein bestimmtes Relationeninventar, Regeln für die Wortarteninterpretation u. a.

Die etablierte Retrievaltechnik hat die Vorteile, robust und voll operationalisierbar zu sein. Beim semantischen Retrieval wird eine intellektuelle Indexierung nötig sein, solange eine automatische Themenzuweisung für Texte noch nicht in einsatzfähigem Zustand entwickelt ist. Die kontinuierlichen Aktivitäten der linguistischen Komplementierung der etablierten Recherchetechnik durch Stammformenbildung, Phrasenextraktion u. a. (Perez-Carballo und Strzalkowski), die Einführung von Sacherschließungsinformation bei Metadaten für Webseiten, die Möglichkeiten, mit XML Dokumente "ontologiebasiert" inhaltlich zu markieren (Rabarijaona et. al.) und anderes deuten aber auf eine Wende hin. Man erkennt, dass man für qualifizierte Information auch einen kleinen intellektuellen Aufwand treiben sollte. Der professionelle Benutzer wägt schließlich ab, ob seine

Rechercheergebnisse in einem angemessenen Verhältnis zu seinem Zeitaufwand und seinen Kosten für die Informationssuche stehen.

Literatur

Klavans, Judith L.: Visions of the Digital Library: Views on Using Computational Linguistics and Semantic Nets in Information Retrieval. Current Issues in Computational Linguistics: In Honour of Don Walker. Ed. By Antonio Zampolli, Nicoletta Calzolari, Martha Palmer. Pisa 1994, p. 227-236.

Lein, Hendrik: Aspekte der Realisierung des semantischen Retrievals. In: Blick Europa! Informations- und Dokumenten-Management. Deutsche Dokumentartag 1994. Trier, 27.-30.9.1994. Herausgegeben von der Deutschen Gesellschaft für Dokumentation, Frankfurt 1994.

Mladenic, Dunja: Text-Learning and Related Intelligent Agents: A Survey. IEEE Intelligent Systems. July/August 1999, p. 44-54

Perez-Carballo, Jose; Tomek Strzalkowski: Natural Language Information Retrieval: Progress Report. In: Information Processing and Management 36 (2000), p. 155-178

Rabarijaona, Auguste; Rose Dieng, Oliver Corby; Rajae Quaddari: Building and Searching an XML-Based Corporate Memory. IEEE Intelligent Systems, May/June 2000, p. 56-63.

Rahmstorf, Gerhard: Use od Semantic Networks for Information Retrieval. In: G. Rahmstorf; M. Ferguson (ed.): Proceedings of a Workshop on Natural Language for Interaction with Data Bases, IIASA (International Institute for Applied Systems Analysis), January 10-14, 1977, Laxenburg 1978.

Rahmstorf, Gerhard: Semantisches Information Retrieval. In: Blick Europa! Informations- und Dokumenten-Management. Deutsche Dokumentartag 1994. Trier, 27.-30.9.1994. Herausgegeben von der Deutschen Gesellschaft für Dokumentation, Frankfurt 1994.

Raphael, Bertram: SIR, A Computer Program for Semantic Information Retrieval. In: Minsky, M. (ed.): Semantic Information Processing. Cambridge, MA, 1968, p. 33-145.

Schank, Roger; Janet Kolodner; Gerald DeJong: Conceptual Information Retrieval. New Haven, Connecticut. Yale University, Research Report 190, December 1980.