Konzeption und prototypische Realisierung einer begriffsbasierten Texterschließung

Dissertation

an der ***Universität Trier*** – Fachbereich IV –

zur Erlangung der Würde eines

Doktors der Wirtschafts- und Sozialwissenschaften

Dr. rer. pol.

vorgelegt von

Diplom-Wirtschaftsinformatiker Sascha Lorenz

Betreuer Univ.-Prof Dr. habil. Hans Czap
Wirtschaftsinformatik I
Universität Trier

Univ.-Prof. Dr. habil. Ralph Bergmann
Wirtschaftsinformatik II
Universität Trier

„Geschrieben steht: ‚Im Anfang war das Wort!‘
Hier stock ich schon! Wer hilft mir weiter fort?
Ich kann das Wort so hoch unmöglich schätzen,
Ich muß es anders übersetzen,
Wenn ich vom Geiste recht erleuchtet bin.
Geschrieben steht: Im Anfang war der Sinn. ...“

Goethe, Faust I

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

IR	Information Retrieval
POS-Tagger	Part-of-Speech-Tagger
TF-IDF	Term Frequency – Inverse Document Frequency
XML	eXtensible Mark-up Language
RDF	Resource Description Format
TL	TrustLevel
GSR	Generalisierungs-Spezialisierungs-Relationen

1 Einleitung

1.1 Problemstellung

Menschen kommunizieren zu einem großen Teil durch ihre Sprache. Durch deren Aufzeichnung als Texte lassen sich Kenntnisse, Fertigkeiten und somit Wissen übertragen. Der maschinelle Zugriff auf das in Texten explizit gemachte Wissen stellt trotz der Unterstützung durch Rechentechnik einen zeit- und damit kostenintensiven Prozess dar, dem große wirtschaftliche Bedeutung zukommt [KöRe98; TsiLa02]. Der sich diesem Problem widmende Forschungszweig des ***Information Retrieval*** (IR) hat eine Vielzahl von Methoden hervorgebracht oder adaptiert, um dieser Herausforderung zu begegnen. Die Problematik der fehlenden inhaltlichen Erschließung mit Hilfe maschineller Verfahren besteht jedoch weiterhin.

Bestehende IR-Systeme zielen auf die Unterstützung des Wiederfindens von Dokumenten oder Teilen davon. Sie orientieren sich an einzelnen, als Indexterme bezeichneten Worten. Diese werden entweder manuell einem Text zugeordnet oder maschinell auf Grund statistischer Maße ermittelt. Eine Suchanfrage liefert dann Verweise auf diejenigen Dokumente, denen diese Indexterme zugeordnet sind. Es werden also keine Informationen über den gesuchten Sachverhalt sondern lediglich Verweise darauf geliefert. Als Konsequenz bleibt es dem Nutzer überlassen festzustellen, ob und in welchem Ausmaß die gesuchten Informationen in den Dokumenten enthalten sind. Der Informationsbedarf wird folglich nur mittelbar befriedigt.

Eine weitere Folge dieser am Wort orientierten Vorgehensweise ist, dass nicht bedeutungsbezogen gesucht werden kann. Da nicht bekannt ist, welchen Begriff ein Term symbolisiert, kann ein IR-System Fragen nach Personen, deren Funktionen, etc. nicht beantworten. Dies wird erst durch die als ***Information Extraction*** bezeichnete, inhaltliche Erschließung möglich. Deren Ziel ist es, unstrukturierte Daten aus Texten so zu strukturieren, dass ein gezielter Zugriff ermöglicht wird.

Neben vielen Ansätzen innerhalb der Information Extraction, die sich sehr speziellen und tiefgehenden Detailfragen widmen, hat sich der Bereich der Erkennung von Eigennamen etabliert. Diese unter der Bezeichnung ***Named Entity Recogni-***

tion zusammengefassten Vorgehensweisen zielen auf besonders häufig anzutreffende Namen wie die von Personen, Organisationen und Orten. Damit lassen sich ähnlich einer Datenbank auch direkte Anfragen nach diesen Eigennamen beantworten, die auf Grund ihrer Seltenheit kaum als Indexterme in Frage kämen. Allerdings zielen diese Verfahren nur auf echte Eigennamen oder durch äußerliche Regelmäßigkeiten eindeutig beschreibbare Größen wie beispielsweise Datumsangaben. Gattungsnamen oder Bezeichnungen abstrakter Sachverhalte werden nicht betrachtet.

Allen diesen Vorgehensweisen gemeinsam ist deren Orientierung am Wort. Sie betrachten einen Text als eine Folge einzelner Worte, die mit einer bestimmten Wahrscheinlichkeit auftreten. Insbesondere die Erkennung von Eigennamen ist daher auf eine vorhergehende syntaktische Analyse der Texte angewiesen, weil dann auf Grund der ermittelten Wortarten auch auf seltenere und damit weniger wahrscheinliche Bezeichner geschlossen werden kann.

Trotz der zusätzlichen syntaktischen Informationen werden große Mengen manuell aufbereiteter Trainingsdaten benötigt, um statistisch relevante Aussagen für maschinelle Lernverfahren treffen zu können. Daher findet die inhaltliche Erschließung im IR praktisch keine Verwendung. Stattdessen werden immer mehr Worte zur Indexierung benutzt, ohne damit jedoch Bedeutungen erschließen zu können.

Der Grund für die Probleme der inhaltlichen Texterschließung ist die historisch gewachsene Orientierung am Wort. Dementsprechend verspricht der Übergang vom Wort zu dem dadurch bezeichneten Begriff ein Weg zur Überwindung dieser Schwierigkeiten zu sein. Basis dieser Überlegungen ist, dass Kommunikation in natürlicher Sprache einem Protokoll folgt, also strukturiert abläuft. Dieses findet sich in Form typischer Kommunikationsmuster auch in der geschriebenen Sprache wieder. Da diese Muster aus zueinander in Beziehung stehenden Begriffen bestehen, erlauben die resultierenden Begriffsnetze die Identifikation gesuchter Bedeutungen sowie die Erschließung begrifflicher Zusammenhänge.

1.2 Zielsetzung

Im Rahmen dieser Arbeit wird eine Vorgehensweise entwickelt, die die Fixierung auf das Wort und die damit verbundenen Schwächen überwindet. Sie gestattet die Extraktion von Informationen anhand der repräsentierten Begriffe und bildet damit die Basis einer inhaltlichen Texterschließung. Die anschließende prototypische Realisierung dient dazu, die Konzeption zu überprüfen sowie ihre Möglichkeiten und Grenzen abzuschätzen und zu bewerten.

Arbeiten zum Information Extraction widmen sich fast ausschließlich dem Englischen, wobei insbesondere im Bereich der Named Entities sehr gute Ergebnisse erzielt werden. Deutlich schlechter sehen die Resultate für weniger regelmäßige Sprachen wie beispielsweise das Deutsche aus. Aus diesem Grund sowie praktischen Erwägungen wie insbesondere der Vertrautheit des Autors damit, soll diese Sprache primär Gegenstand der Untersuchungen sein.

Die Lösung von einer engen Termorientierung bei gleichzeitiger Betonung der repräsentierten Begriffe legt nahe, dass nicht nur die verwendeten Worte sekundär werden sondern auch die verwendete Sprache. Um den Rahmen dieser Arbeit nicht zu sprengen wird bei der Untersuchung dieses Punktes das Augenmerk vor allem auf die mit unterschiedlichen Sprachen verbundenen Schwierigkeiten und Besonderheiten gelegt.

1.3 Vorgehensweise

Die Grundlage schriftlicher Kommunikation ist das Wort. Daher wird zu Beginn der Zusammenhang zwischen Wort und Begriff thematisiert. Darauf aufbauend ist die Frage zu untersuchen, welchen Einfluss das zu Grunde liegende Begriffsverständnis auf das Information Retrieval hat und ob sich daraus ein neuartiger Ansatz ableiten lässt.

Da geschriebene Sprache der Externalisierung und Übertragung von Wissen dient, ist auch zu klären, was in diesem Zusammenhang unter Wissen zu verstehen ist und wie es sich sprachlich äußert. Abgerundet wird dieser erste Teil durch die Erörterung, was Sprache ist und wie sie maschinell verarbeitet werden kann.

1.3 Vorgehensweise

Daran schließt sich eine Darstellung wichtiger Prinzipien und Verfahren des IR an. Dies ist notwendig, um innerhalb der traditionellen Vorgehensweise einerseits Gründe für Unzulänglichkeiten und andererseits Bewährtes zu identifizieren. Außerdem ist zu erwarten, dass sich hieraus weitere wichtige Aspekte offenbaren, die bei der verfolgten, am Begriff orientierten Vorgehensweise direkt zu berücksichtigen sind oder Potentiale für spätere Verbesserungen bergen.

Die Frage der Repräsentation von Informationen schlägt dann den Bogen zwischen wort- und begriffsorientiertem Ansatz und bildet die Überleitung zur inhaltlichen Erschließung sowie zur Informationsextraktion. Mit der aus den dargestellten Grundlagen abgeleiteten Begründung einer an der Kommunikation orientierten Herangehensweise schließt Kapitel 2 ab.

Das dritte Kapitel widmet sich der Entwicklung einer Vorgehensweise zur inhaltlichen Erschließung von Texten in natürlicher Sprache. Aufbauend auf die Darstellungen im Grundlagenteil wird dazu ein Konzept entworfen, welches letztlich auf ein an der Verwendung in der Sprache orientiertes Begriffsverständnis aufbaut. Ein wesentlicher Aspekt ist dabei die Identifikation von Kommunikationsmustern, die ihrerseits als für eine Sprache typische Konstellationen von Begriffen aufgefasst werden.

Nachdem diese Vorgehensweise prinzipiell dargestellt worden ist, erfolgt deren prototypische Umsetzung. Neben der Verifikation der grundlegenden Eignung des Konzepts soll damit vor allem untersucht werden, unter welchen Voraussetzungen welche Ergebnisse zu erwarten beziehungsweise welche Anpassungen nötig sind. Durch die im Kapitel 3 verfolgte Top-Down-Vorgehensweise von der Konzeption über die Detaillierung bis zur Umsetzung soll eine gute Übertragbarkeit auf neue Gegebenheiten erreicht werden.

Mit der Evaluation der in der Realisierung erzielten Ergebnisse und damit der zu Grunde liegenden Vorgehensweise findet die Arbeit in der Diskussion ihren Abschluss.

2 Grundlagen

Im Rahmen dieses Kapitels der Arbeit soll zunächst ein Überblick über die verschiedenen Sichtweisen auf den Zusammenhang von Bezeichnung, Sachverhalt und Begriff sowie Wissen gegeben werden. Danach erfolgt eine Übersicht der Aspekte von Sprache sowie Sprachverarbeitung.

Im Selbstverständnis dieser Arbeit umfasst Information Retrieval nicht nur das Auffinden von Informationen in Textdokumenten. Vielmehr sollen darunter alle Maßnahmen verstanden werden, um gezielt auf Informationen zugreifen zu können. Daher erfolgt nach der Darstellung von Methoden des herkömmlichen IR sowie darauf aufbauender Erweiterungen die Überleitung zum Bereich der Informations- und Wissensrepräsentation. Davon ausgehend richtet sich der Fokus auf die inhaltliche und begriffliche Erschließung von Informationen aus Texten.

2.1 Bezeichnung, Begriff und Wissen

Das einleitende Zitat des zweiten Deckblatts aus Goethes ‚Faust' [Goet08, S.42f] ist nur ein Beleg dafür, dass sich Menschen schon seit sehr langer Zeit mit der Frage nach der Bedeutung der Worte beschäftigen[1]. Eine erste Systematisierung des Verhältnisses von Wort und Sinn geht zurück auf Aristoteles [Arist07, S.124], der in seiner Metaphysik einen ***indirekten*** Zusammenhang zwischen einem Sachverhalt und seiner Bezeichnung postulierte [Boni90].

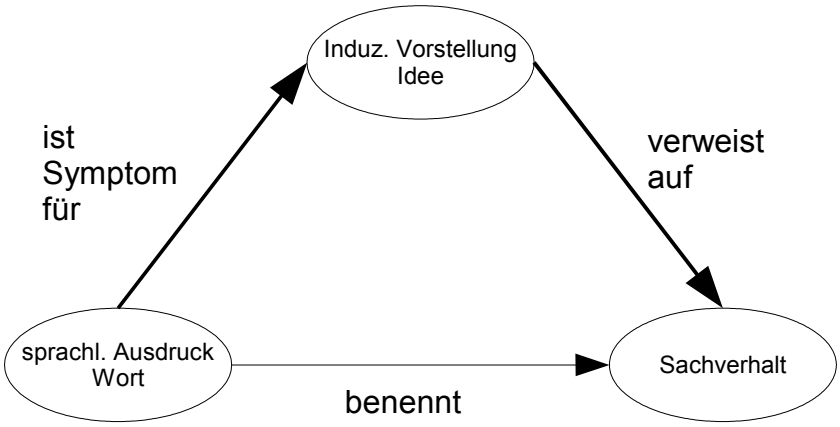
2.1.1 Sichtweisen auf Begriffe

Nach Aristoteles ruft ein Sachverhalt eine ‚Seelenregung', also eine Vorstellung von diesem Sachverhalt beim Menschen hervor, die mit Hilfe eines Bezeichners zum Ausdruck gebracht wird. Dieser Zusammenhang wird als semiotisches Dreieck bezeichnet, wie in Abbildung 1 dargestellt. Die Verbindung zwischen dem Wort und dem Sachverhalt ist demnach durch die Vorstellung der Person gegeben. Diese ordnet einem Sachverhalt einen sprachlichen Ausdruck, also beispielsweise ein Wort, zu, der je nach Person und Kontext unterschiedlich ausfallen kann [Gric77].

1 Ein weiteres Beispiel für die Trennung zwischen Bezeichnung und Bedeutung findet sich in der Genesis: „Er formte aus dem Erdboden die Landtiere und die Vögel und brachte sie zu dem Menschen, um zu sehen, wie er sie nennen würde. Genauso sollten sie dann heißen."[Bibel, S.2]

Diese, an sich willkürliche Verbindung zwischen dem Wort und dem beobachteten Phänomen bildet den individuellen Begriff, die Vorstellung der Person von diesem Sachverhalt [Lyons75, S.55f]. Der Begriff an sich ist demnach abstrakt [Volk79, S.48f] – erst durch Verknüpfung mit dem Erfahrungswissen des Rezipienten erhält ein Wort eine konkrete Bedeutung und wird zur Repräsentation eines Begriffs im üblichen Sinn [Bühl65; Staab02]. Teilen mehrere Personen einen gemeinsamen Erfahrungshintergrund, werden deren Begriffe gemeinsam betrachteter Sachverhalte jeweils sehr ähnlich sein. Dieses gemeinsame Begriffsverständnis äußert sich im spezifischen Vokabular fachlicher Gruppen [Trier73a].

Abbildung 1: Das semiotische Dreieck nach Aristoteles



Quelle: eigene Erstellung in Anlehnung an [Boni90]

2.1.1.1 Extensionale und intensionale Sichtweise

Ausgehend von diesen Überlegungen leitet sich im einfachsten Fall ein extensionales Begriffsverständnis ab, demzufolge ein Begriff durch die Summe der bezeichneten Objekte bzw. deren Bezeichnungen definiert ist. So könnte beispielsweise der Begriff ‚Frieden' durch die Menge der Zeichenfolgen {‚Idyll', ‚Ruhe', ‚Frieden', ‚Freundschaft', ‚kein Krieg'} beschrieben sein. Vor dem Hintergrund der philosophischen Orientierung Aristoteles ist es nicht verwunderlich, dass diese Sichtweise vor allem für abstrakte und zusammenfassende[2] Begriffe geeignet ist.

Die extensionale Auffassung bildet bei genauerer Betrachtung eine wesentliche Basis der im Kapitel 2.2, Information Retrieval, dargestellten indexorientierten Vorgehensweise beim IR. Auch hier wird eine Menge von Worten – der Index – dazu

2 z.B.: Werkzeug={Zange, Hammer, Meißel, Bohrer,...}

benutzt, einen Sachverhalt, hier die in einem Dokument enthaltenen Informationen, zu beschreiben. Entsprechend wird davon ausgegangen, dass größere Mengen von Worten eine genauere Beschreibung der Inhalte ermöglichen.

Die intensionale Sichtweise auf einen Begriff basiert auf der Angabe der spezifischen Merkmale des bezeichneten Sachverhalts [Volk79, S.37ff]. Damit grenzen sich Begriffe gegeneinander durch Ausdifferenzierung anhand gemeinsamer und unterscheidender Merkmale ab [Czap96, S.4]. Smith und Medin [SmMe81] unterscheiden weiter in eine probabilistische Sicht nach Mill [Mill65, S.277] sowie die prototypische Sicht nach Whewell [Whew58, S.122].

Dabei betrachtet die probabilistische Sicht Mengen von Merkmalen als begriffsbildend, während aus prototypischer Sicht Konzepte anhand übereinstimmender Merkmale klassifiziert werden [Sowa84, S.16f]. Letztere geht also von einem exemplarischen Begriff aus und beschreibt diesen durch eine Menge von Merkmalen. Weist nach probabilistischer Sicht ein Sachverhalt eine Mindestmenge bestimmter, als essentiell betrachteter Merkmale auf, so handelt es sich um eine Instanz dieses Begriffs [GaWi99]. In Summa bezeichnet die intensionale Sicht Begriffe als ähnlich oder identisch, wenn die zu Grunde liegenden Sachverhalte in ihren wesentlichen Merkmalen weitgehend oder vollständig übereinstimmen.

Die unterschiedlichen, jedoch eng verwandten Herangehensweisen nach probabilistischer und prototypischer Sicht spiegeln sich in der Unterscheidung in Klassifikation und Clustering im Rahmen der Datenanalyse wieder (siehe auch Abschnitt 2.4.2 ab S.66). Anhand von Abstufungen innerhalb der Merkmalsmengen entsprechend des Detaillierungsgrades sind Unterscheidungen im Sinne von Taxonomien, also hierarchisch angeordneten Klassen, möglich.

Aus der intensionalen Sichtweise leitet sich die oft anzutreffende Gleichsetzung von Begriff und Bedeutung ab [Bußm02, S.128]. Diese Sicht ist sehr typisch für die Erfassung von Informationen in Datenbanken. Legt man das relationale Datenmodell zu Grunde, sind Sachverhalte Entitäten mit ihren Attributen und Werten. Ähnlichkeit definiert sich damit über die Anzahl gemeinsamer Merkmalsausprägungen bei den als relevant betrachteten Attributen. Im Rahmen der formalen Begriffsanalyse wird daraus auf semantische Zusammenhänge geschlossen [Ga-

Wi99]. Eine Übertragung auf sprachliche Probleme zeigt Ganter [Gant04, S.72ff]. Der mit dieser Strukturierung aus technischer Sicht verbundene Vorteil ist die Möglichkeit des gezielten Zugriffs auf die Informationen.

Die Menge der tatsächlich verwendeten sprachlichen Konstrukte zur Bezeichnung eines Sachverhalts spiegelt die Summe der individuellen Sichten auf einen Sachverhalt wieder [Lyons75, S.55f]. Deutlicher wird diese Auffassung in der englischen Übersetzung des Wortes ‚Begriff'. Hier heißt es ***concept*** oder ***notion***, was wiederum soviel heißt wie Konzept, Plan oder Gedanke, Idee. Damit wird der semantische Aspekt, wie ihn auch das semiotische Dreieck (Abbildung 1) suggeriert, deutlich.

Verschiedene Personen interpretieren Begriffe auch verschieden bzw. verstehen unterschiedliche Dinge darunter. Deutlich wird diese individuelle Interpretation z.B. in den Aspekten einer Botschaft nach Schulz von Thun [Schul81][3]. Bezogen auf die extensionale Herangehensweise spiegelt sich dies in der Vielzahl von Bezeichnungen für einen Sachverhalt (Homonyme) wider, während aus intensionaler Sicht unterschiedliche Merkmale bzw. Merkmalsausprägungen individuell als wesentlich für die Zuordnung von Sachverhalt und Begriff angesehen werden. Daraus wird deutlich, dass zur Durchführung von Kommunikation eine gemeinsame Begriffs- und Bezeichnungswelt nötig ist, um Missverständnisse zu vermeiden [Czap88].

Hier setzt auch die Kritik an der intensionalen Auffassung an. Es ist nicht möglich, einen Begriff durch Aufzählung seiner Merkmale vollständig und damit eindeutig zu beschreiben [Witt53], da die Auswahl der Merkmale immer aus einer speziellen Sicht heraus geschieht und sich nicht weiter zerlegbare, atomare Merkmale nur selten finden lassen [Witt18]. Sowa illustriert dies am Beispiel eines Stuhls: Für die meisten Menschen ist es einfach eine Sitzgelegenheit; für einen Schreiner eine Vielzahl von Teilen, die sorgfältig zusammenzufügen sind [Sowa84, S.15].

Zur Durchführung sprachlicher Kommunikation ist es nötig, eine gemeinsame Vorstellung der betrachteten Sachverhalte zu finden, was wiederum in einem gemeinsamen Vokabular zum Ausdruck kommt [Grub93, S.2; Dahl74, S.111f]. Da nicht für alle denkbaren Entitäten vorab Bezeichner vereinbart sind, muss sich deren Bedeu-

3 Schulz von Thun beschäftigt sich dabei mit der Frage, wie Missverständnisse entstehen und vermieden werden können und zeigt, welchen Einfluss die Interpretation von Signalen dabei hat.

tung aus dem Zusammenhang ergeben. Damit bildet die Menge der sprachlichen Äußerungen über einen Sachverhalt einen Ansatz zur Erschließung der Bedeutung eines Bezeichners.

2.1.1.2 Kontextorientierte Sichtweise

Aus der Sprachphilisophie nach Wittgenstein [Witt18] und Bühler [Bühl65] leitet sich eine Sichtweise ab, die den Begriff durch sein Umfeld, also die Verwendung in der Sprache definiert[4]. Diese Sichtweise sei hier als Kontextsicht bezeichnet. Sprache ist gemäß dieser auch als Organonmodell bezeichneten Sicht einfach ein Werkzeug[5] zur Kommunikation. Der folgende Satz verdeutlicht das Prinzip am Beispiel:

> *„Kartellrechtliche Probleme sehen Adobes CEO Bruce Chizen und Macromedias Vorsitzender Stephen Elop aber nicht."*
>
> *[HON05a]*

Obwohl nur den wenigsten Personen die genannten Namen vertraut sein werden, ist dem Leser klar, dass es sich dabei um Personen in bestimmten Positionen innerhalb bestimmter Unternehmen handelt. Der Grund dafür liegt in unserem Verständnis von Sprache, also unserem Begriff davon, wie Sprache funktioniert. Dieses Sprach- und Begriffsverständnis findet sich auch in der Wortfeldtheorie nach Jost Trier [Trier73] und John Lyons [Lyons80] wieder. Das Deutsche Institut für Normung definiert in DIN 2342 einen Begriff rekursiv als „Begriffsbestimmung mit sprachlichen Mitteln" [Herz06]. Dies wird im Allgemeinen als Summe aller Aussagen bezüglich des bezeichneten Sachverhalts gesehen.

Aufbauend auf der Kritik an den anderen Begriffsauffassungen und den philosophischen Überlegungen Wittgensteins und anderer entwickelte Sowa ein Modell zur Abbildung begrifflicher Zusammenhänge [Sowa84, S.69ff], welches in Abschnitt 2.3.2.2 ab Seite 58 vorgestellt wird.

Nach Lyons ermöglicht die am Kontext bzw. der Verwendung orientierte Sichtweise das Erlernen von Sprache nachzuvollziehen [Lyons75, S.417, S.419f]. Um einem Gedanken bzw. Sachverhalt Ausdruck zu verleihen, spricht man vom ihm so, dass

4 Insbesondere sei auf die Sektionen 2.0ff und 2.1ff in [Witt18] hingewiesen.
5 Organon – (griechisch) = Hilfsmittel, Werkzeug

beim Empfänger dieselbe Vorstellung wie beim Sender hervorgerufen wird – die Wahl der Worte ist dabei zweitrangig [FreMo72, S.13]. Allerdings bringt die damit verbundene Flexibilität auch enorme Schwierigkeiten, wenn man darauf aufbauende Informationssysteme realisieren will. In letzter Konsequenz erfordert diese Sichtweise ein Erlernen der Sprache.

2.1.2 Konsequenzen für die Arbeit

Unabhängig von der verwendeten Auffassung über das Wesen eines Begriffs ist festzustellen, dass die Beurteilung der Zulässigkeit einer Begriffsbezeichnung für einen Begriff eine individuelle Entscheidung darstellt, die nicht zwangsweise mit der anderer Personen übereinstimmt [Wil63, S.54]. Da Kommunikation jedoch ein gemeinsames Begriffsverständnis voraussetzt, soll nur dieser Fall betrachtet werden. Übertragen auf textuelle Kommunikation bedeutet dies eine Einschränkung des jeweils zu betrachtenden Diskursbereichs. Desto eingegrenzter dieser ist, desto größer ist die begriffliche Übereinstimmung der an der Kommunikation beteiligten Gruppen. Damit wird die Art und Weise, in der Begriffe in Beziehung stehen konsistent sein und sich in typischen, rekonstruierbaren Mustern äußern [Onto04; Quil85].

Die Zuordnung eines Bezeichners zu einem Begriff entspricht aus technischer Sicht einer Kategorisierung oder Klassifikation [Dahl74, S.115f]. Dies bedeutet, dass die Aufgabe eines begriffsbasierten Information Retrievals darin besteht, Zeichenfolgen als gültige Bezeichnungen von Begriffen zu identifizieren und damit zu klassifizieren. Auch auf Grund der Individualität der als richtig empfundenen Klassifikationsentscheidungen besteht die Notwendigkeit einer thematischen Eingrenzung der in Frage kommenden Textdaten. Dennoch ist es außer bei Faktenentscheidungen nicht möglich, eine Klassifikation allgemein gültig als richtig zu bezeichnen, was die Beurteilung von Klassifikationen erschwert [Dahl74, S.112]. Wilson erörtert diese Problematik anhand der auf den ersten Blick einfachen Frage, ob ein Wal ein Fisch sei [Wil63, S.7]. Wurden diese sprachlichen Konventionen nicht explizit vereinbart, kann die Gültigkeit einer Klassifikation innerhalb des Diskursbereichs[6] nur durch die Häufigkeit ihres Auftretens abgeschätzt werden [Schu95, S.75].

6 Die Gültigkeit einer Klassifikation kann auch auf eine kommunizierende Gruppe innerhalb einer Diskursbereichs eingeschränkt sein.

Fügt man diese beiden Aspekte zusammen, so ergibt sich die Überlegung, mit Hilfe der Kommunikationsmuster die Klassifikation der damit kommunizierten Begriffe durchzuführen. Wird ein Kommunikationsmuster typischer Weise dazu benutzt, die Information über den Namen (=Bezeichner) einer Person (=Begriff) weiter zu geben, so kann dieses Muster genutzt werden, weitere Bezeichner für Instanzen des Begriffs Person maschinell zu erschließen.

Desto größer die thematische Eingrenzung des zu Grunde liegenden Textmaterials ist, desto sicherer wird die Klassifikation sein, da weniger individuell verschiedene Vorstellungen über die beschriebenen Sachverhalte bestehen. Auf Grund der einseitigen Kommunikation, die keine Nachfragen ermöglicht, müssen Nachrichten klar formuliert sein, also typischen Kommunikationsmustern folgen. Weiterhin haben Nachrichten einen zu vermittelnden Inhalt, was ebenfalls zu einer thematischen Einschränkung führt. Damit gewährleistet dieser Texttyp am ehesten, dass Kommunikationsmuster hinreichend häufig auftreten und so gute Klassifikationsergebnisse erzielt werden können. Aus diesen Gründen soll sich die prototypische Realisation auf Korpora von Nachrichten konzentrieren.

Um die weitere Bezeichnungsweise zu systematisieren, soll folgende Taxonomie in Anlehnung an [Arist07, S.123 und 297], wie sie auch im Duden vorgeschlagen wird, Verwendung finden: Handelt es sich bei einem Sachverhalt um eine Menge unterscheidbarer Elemente, so wird der Sachverhalt als Klasse oder Gattung bezeichnet und dieser Klasse ein Begriff als Gattungsname zugeordnet. Ein anonymes Element dieser Menge wird als Objekt bezeichnet, während ein konkretes Element als Instanz oder Entität bezeichnet wird. ‚Klaus' ist demnach die Bezeichnung bzw. der Eigenname einer Instanz der durch den Begriff ‚Menschen' bezeichneten Klasse. Mit ‚Mann' werden hingegen im Beispiel ein oder mehrere anonyme Elemente (Objekte) der Klasse bezeichnet.

Ein Begriff impliziert also stets eine bestimmte Abstraktionsebene. Interessieren nicht die konkreten Instanzen sondern nur Teilmengen der Klasse, so werden diese jeweils mit einem Begriff bezeichnet. Bei abstrakten, nicht beobachtbaren Sachverhalten werden entsprechend auch individuelle Elemente durch Begriffe bezeichnet. Beispielsweise ist der Begriff ‚Friede von Ketomar' eine Instanz der Klasse histo-

rischer Friedensschlüsse. Wird im Rahmen dieser Arbeit von abstrakten Begriffen gesprochen, so meint dies in erster Linie Konzepte, die zur Strukturierung und Erklärung eines Diskursbereichs genutzt werden. Sie setzen Instanzen zueinander in Beziehung [Sowa84, S.76]. Ein Beispiel dafür ist der Begriff der ROLLE, also der Funktion die eine Person in einer Organisation ausübt.

Nachdem aus dem Begriffsverständnis die Idee entwickelt wurde, Kommunikationsmuster zur Klassifikation von Begriffsbezeichnern zu benutzen, stellt sich die Frage, wie diese Muster aus den natürlich sprachigen Texten gewonnen und abgebildet werden können. Dies wird in Abschnitt 2.1.4 erörtert. Da im Zusammenhang mit Begriffen und Informationen oft auch von Wissen die Rede ist [Czap96, S.4; Grub93; Nohr00], soll zuvor dieser Aspekt, soweit notwendig, untersucht werden.

2.1.3 Wissen

2.1.3.1 Traditionelle Sichtweisen

In den verschiedenen Zweigen der Forschung existieren unterschiedliche Sichten auf Wissen, die in erster Linie durch unterschiedliche Anwendungsschwerpunkte bedingt sind [FrSc04]. Die folgende Zusammenfassung der drei wichtigsten Sichten orientiert sich an [ReiMa02].

Die behavioristische Sicht stellt im Wesentlichen auf die Verbindungen der Nervenzellen im Gehirn ab. Daher werden vor allem die Reiz-Reaktion-Verbindungen untersucht, denen die Speicherung von Aktivitätsmustern unterstellt werden. Auf dieser Wissenssicht beruht der Konnektionismus im Sinne künstlicher neuronaler Netze sowie darauf aufbauend die Assoziationstheorie. Hier sei auf die in Kapitel 2.4.3 dargestellten konnektionistischen Verfahren verwiesen.

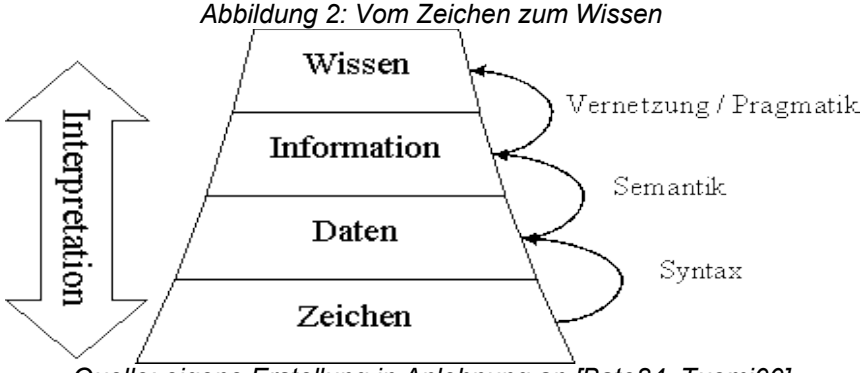
Aus kognitiver Sicht wird Wissen als der Besitz von Konzepten, also Vorstellungen und Ideen, sowie kognitiven Fähigkeiten gesehen, die die Wiedererkennung und Konstruktion von Symbolmustern erlauben. Diese konstruktivistische Sichtweise unterstützt die These des individuellen, kontextabhängigen Wissens, welches durch die Verknüpfung von Informationen durch ein Individuum entsteht. Als Kerngedanke ergibt sich die Fähigkeit zur Abstraktion und Beurteilung von Ähnlichkeit, was seinen Niederschlag beispielsweise im ***Case Based Reasoning*** findet [Czap96, S.5].

Die situative Auffassung von Wissen greift eine sehr frühe, bereits bei Descartes [Desc28a] zu findende Vorstellung wieder auf, nach der Wissen in der Welt verteilt und somit per se vorhanden ist. In diesem Zusammenhang wird oft das Zitat „Alles Wissen besteht in einer sicheren und klaren Erkenntnis.“ [Desc28] gebracht. Eine der ältesten Arbeiten dieses Tenors ist das ‚Novum Organum‘ von Francis Bacon [Baco20, Liber Primus, Satz 8]. Diese pragmatische Sicht stellt somit die Erkenntnis über die vorhandenen Zusammenhänge mit der Erlangung von Wissen gleich [Desc28]. Obwohl diese Auffassung sehr unspezifisch wirkt, trifft sie doch die übliche Auffassung vom Wissen als Wissen über Sachverhalte, dem Faktenwissen.

2.1.3.2 Technische Sicht

Aus technischer Sicht steht der Zusammenhang zwischen Informationen im Mittelpunkt, wobei unter Wissen oft der Zusammenhang an sich und unter Intelligenz die Nutzung dieses Zusammenhangs mit Hilfe logischer Schlüsse und daraus resultierendes Handeln verstanden wird [Hoff01].

Abbildung 2: Vom Zeichen zum Wissen



Quelle: eigene Erstellung in Anlehnung an [Bate84, Tuomi00]

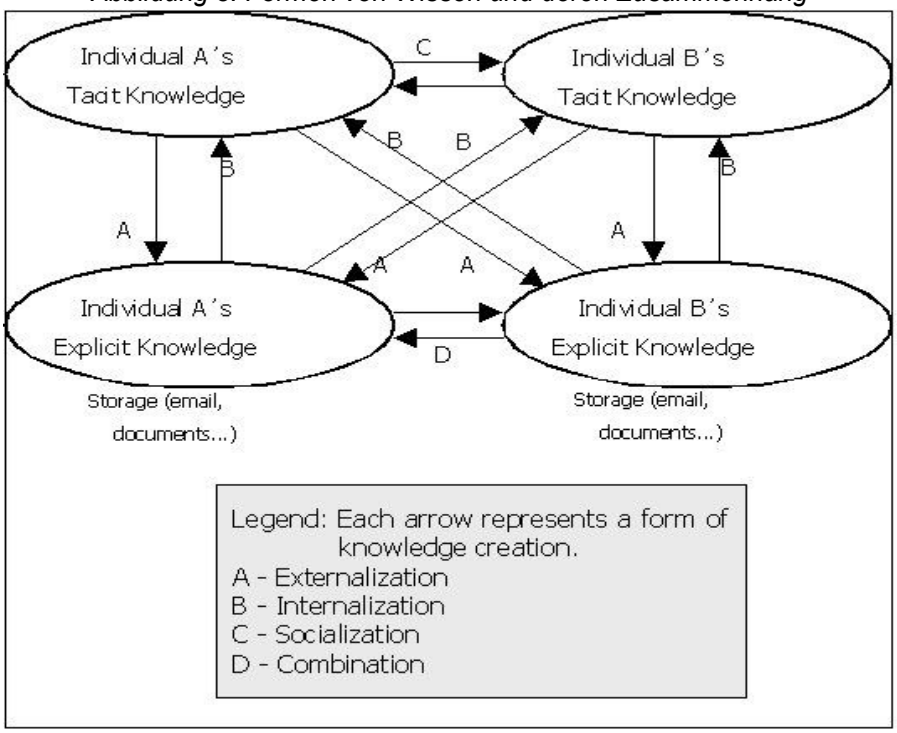
Aus der Semiotik stammt eine häufig anzutreffende Einteilung in Zeichen, Syntax und Semantik, die darauf aufbauend auch Daten, Informationen und Wissen mit einbezieht. Abbildung 2 veranschaulicht dies, wobei darauf hinzuweisen ist, dass das bloße Vernetzen von Informationen nicht per se Wissen schafft. Erst mit der Aufnahme der Zusammenhänge und deren Interpretation durch ein Individuum als Wissensträger kann sinnvoll von Wissen gesprochen werden [Tuomi00, KarTe01, S.20]. Nonaka und Takeuchi [NoTa95] sprechen hier auch von ‚tacit knowledge’, also „expressed or carried on without words or speech”[7].

7 Nach Merriam-Webster Online Dictionary; Abruf am 2006-03-08

2.1.3 Wissen

Wissen äußert sich dann in der Fähigkeit eines Menschen, eine bestimmte Aufgabe zu erfüllen [Staab02]. Diese Form wird als Handhabungs- bzw. Orientierungswissen bezeichnet, während Faktenwissen auf die Kenntnis über Sachverhalte abzielt [Czap96]. Beide Aspekte sind jedoch zur Aufgabenerfüllung nötig. In Abbildung 3 wird die Bedeutung der Aufnahme durch ein Individuum illustriert.

Abbildung 3: Formen von Wissen und deren Zusammenhang



Quelle: [AlLe01, S.117] in Anlehnung an [NoTa95]

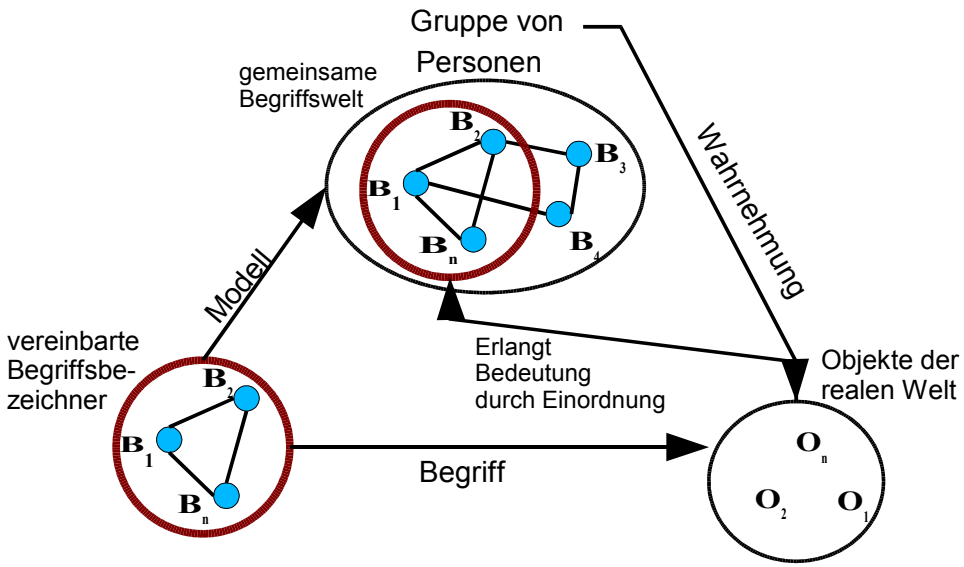
2.1.3.3 Verbindung zur Arbeit

Während Handlungs- oder Orientierungswissen ein Untersuchungsgegenstand der künstlichen Intelligenz sind, spielt aus Sicht des Information Retrieval das Faktenwissen die zentrale Rolle. Die Erschließung der Bedeutung von Bezeichnern kann interpretiert werden als Identifikation von Wissenseinheiten. Eine solche Einheit besteht dabei aus dem Bezeichner sowie dem dadurch repräsentierten Begriff. Auf Grund der Klassifikation können die Wissenseinheiten aufgenommen und in Faktenwissen überführt werden.

Vor diesem Hintergrund wird deutlich, warum der Zusammenhang, also die Verbindung zwischen den Wissenseinheiten eine wichtige Rolle spielt. Erst durch Einordnung und Vernetzung von Sachverhalten entsteht Verständnis [Pole78, S.164].

Diesen Aspekt verdeutlicht Abbildung 4 in Form eines erweiterten semiotischen Dreiecks. Tritt ein Begriff B_n im Kontext mit B_1 und B_2 auf, bezeichnet B_n das reale Objekt O_n, welches im Zusammenhang mit den Objekten O_1 und O_2 steht. Bedeutung erlangt das wahrgenommene Objekt durch Einordnung in die Begriffswelt, was sich wiederum in der Bezeichnungssystematik widerspiegelt.

Abbildung 4: Kontextorientierte Erweiterung des semiotischen Dreiecks



Quelle: eigene Erstellung

Die Zuordnung des Bezeichners ‚Hund' zu einem Tier entspricht der symbolischen Repräsentation und gibt das Faktenwissen wieder. Erst durch Einordnung in vorhandenes Wissen wie ‚Hunde sind Säugetiere', ‚Bello ist ein Hund', ‚Klaus hat einen Hund', etc. erschließt sich der Begriff und wird die zu Grunde liegende Begriffswelt deutlich.

Das bedeutet wiederum, dass sprachliches bzw. begriffliches Wissen als Kenntnis über die Beziehungen von Begriffen zueinander gesehen werden kann [FrSc04]. Damit wäre ein Aspekt der Wissensgewinnung die Erlangung von Kenntnissen über den Zusammenhang von Begriffen. Innerhalb eines Diskursbereichs ist damit im obigen Beispiel ein Begriff ‚Hund' definiert. Für andere Diskursbereiche, beispielsweise der Veterinärbiologie, wäre dieser Kontext unpassend.

Die effiziente Aufnahme der gespeicherten Informationen zur Aneignung von Wissen setzt geeignete Mechanismen voraus, so dass ein gezielter Zugriff möglich ist [PuSt+03]. Dabei sollen einerseits vorhandene, aber für den konkreten Wissensbedarf nicht relevante Informationen außen vor bleiben und andererseits auch keine relevanten Informationen fehlen. Dieser selektive Zugriff wird als Retrieval bezeichnet und ist Gegenstand des Abschnitts 2.2.

2.1.4 Sprachen und Sprachverarbeitung

Die folgenden Abschnitte widmen sich zuerst der einfachsten Klasse von Sprachen als Grundlage aller anderen Sprachformen. In Vorbereitung des das Thema Sprache abschließenden Teils Computerlinguistik wird dann auf kontextsensitive Sprachen näher eingegangen. Zu sonstigen Sprachen, deren Ordnung untereinander und Automaten sei auf [Part+90, S.451ff], [Lang04] sowie [Chom95] verwiesen, da hier nur auf die für das prinzipielle Verständnis notwendigen Grundlagen eingegangen werden kann.

Allen Sprachen gemein ist, dass sie auf Gruppen von Zeichen (Worten) eines begrenzten Zeichenvorrats, dem Alphabet, bestehen. Zunächst wird in Anlehnung an [Lang04] der Begriff der Sprache definiert:

Definition I: Sprache

> *Sei A ein Alphabet. Mit A* bezeichnet man die Menge aller Wörter über A. Jede Teilmenge L von A* heißt Sprache über A.*

Bezüglich der Mächtigkeit lassen sich Sprachen in endliche und unendliche Mengen unterteilen. Endliche Sprachen werden durch Aufzählung ihrer Worte angegeben. Im klassischen IR anzutreffende Indexierungssprachen sind ein typischer Vertreter dieser Art, da nur fest vorgegebene Worte zur Beschreibung der Inhalte verwendet werden können. Um unendliche Sprachen angeben zu können, benötigt man eine endliche Beschreibung dieser Sprache. Folglich muss diese Beschreibung Regeln beinhalten, die die Bildung der Worte festlegen, so dass entschieden werden kann, ob ein gegebenes Wort zu der Sprache gehört oder nicht. Die Menge dieser Regeln wird als Syntax der Sprache bezeichnet.

2.1.4.1 Reguläre Sprachen

Reguläre Sprachen sind die einfachste Form von Sprachen und wie folgt definiert:

Definition II: Reguläre Sprache, regulärer Ausdruck

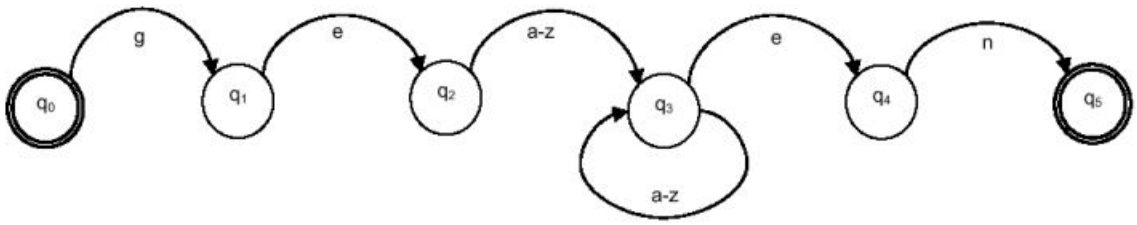
> *Ein regulärer Ausdruck x ist eine Formel zur Erzeugung einer gewissen Teilmenge von A*, der regulären Sprache L(x).*

Reguläre Sprachen sind also durch einen regulären Ausdruck (***Regular Expression***) beschreibbar, wobei dieser Ausdruck aus der disjunktiven Verknüpfung weiterer Ausdrücke bestehen kann. Einen Überblick über die Syntax regulärer Ausdrücke gibt Tabelle 8 auf Seite 92.

Beispiel: Sei *A* das Alphabet der deutschen Sprache und der reguläre Ausdruck laute *x = ge[a-z]+en*. Die Zeichenfolge ‚[a-z]' symbolisiert die Klasse der Kleinbuchstaben, aus der wegen des Modifikators ‚+' mindestens eines auftreten muss.

Damit beschreibt dieser Ausdruck alle Worte, die mit der Zeichenfolge ‚ge' beginnen, mindestens einen weiteren Kleinbuchstaben enthalten und mit der Zeichenfolge ‚en' enden. Dazu gehören beispielsweise ‚gebrauchen', ‚gelaufen', ‚geben' und ‚gezzzzzzen'. Zu jeder Sprache lässt sich ein endlicher Automat angeben, der diese Sprache generieren bzw. akzeptieren kann und daher auch als Akzeptor bezeichnet wird [Part+90; Lang04].

Abbildung 5: Endlicher Automat zum Beispiel



Quelle: eigene Erstellung

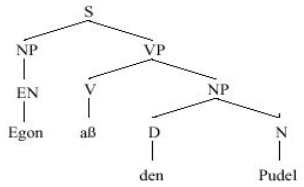
Für den obigen Beispielausdruck gibt Abbildung 5 den entsprechenden Zustandsgraph des Automaten wieder. Damit wird deutlich, dass die durch einen regulären Ausdruck erzeugte Sprache eindeutig bestimmt ist. Eine Sprache kann durch verschiedene reguläre Ausdrücke beschrieben werden, die sich jedoch ineinander überführen lassen.

2.1.4 Sprachen und Sprachverarbeitung

Die Möglichkeiten regulärer Ausdrücke sind, bezogen auf die natürliche Sprache, sehr eingeschränkt. Aus linguistischer und speziell computerlinguistischer Sicht legt man daher eine andere Herangehensweise an die Erkennung von Sprache und damit eine andere Klasse von Sprachen zu Grunde.

Eine weitere, äquivalente Beschreibung einer Sprache ist ihre Grammatik. Aus formaler Sicht ist eine durch eine Grammatik G beschriebene Sprache ein Quadrupel G = (N, T, P, S), wobei N Nonterminale, T Terminale, P Produktionen sind und S das Startsymbol ist. Lässt sich ein Satz auf das Startsymbol zurückführen, so handelt es sich um einen Satz der beschriebenen Sprache (vgl. Abbildung 6 oben).

Abbildung 6: Kontextfreie Grammatik und Syntaxbaum



Beispiel: gegeben folgende (kontextfreie) Grammatik

$$G_1 = \langle \{S, NP, EN, VP, V, D, N\}, \{Egon, Pudel, den, ass\}, R, S \rangle$$

mit der Regelmenge R =

$$\begin{array}{lll} S \rightarrow NP\,VP & NP \rightarrow EN & NP \rightarrow D\,N \\ VP \rightarrow V\,NP & EN \rightarrow \text{Egon} & N \rightarrow \text{Pudel} \\ V \rightarrow \text{ass} & D \rightarrow \text{den} & \end{array}$$

Quelle: [Klen05, S.18]

Unter Terminalen versteht man das Vokabular einer Sprache, also die zugehörigen Worte. Nonterminale entsprechen den grammatischen Strukturen, die ihrerseits aus mehreren Terminalen oder Nonterminalen bestehen. Es gibt verschiedene Arten von Nonterminalen wie beispielsweise Nominal- oder Verbalphrasen, die letztlich die Terminale in bestimmte, für die Sprache typische Gruppen zusammenfassen. Die Produktionen oder Regeln sind die Gesamtheit der Vorschriften, die angeben, welche Nonterminale aufeinander und welche Terminale welchen Nonterminalen folgen dürfen. Anhand eines Beispiels wird dies in Abbildung 6[8] verdeutlicht.

8 Aus [Klen05] einschließlich Rechtschreibfehler übernommen.

Entsprechend unterschiedlicher Möglichkeiten der Grammatiken zur Realisierung komplexer Produktionen mit Rückbezügen werden verschiedene Typen von Grammatiken unterschieden. Reguläre Sprachen werden als Typ-3-Grammatiken bezeichnet und bilden die unterste Ebene der sogenannten Chomsky-Hierarchie [Chom02].

2.1.4.2 Kontextfreie und -sensitive Sprachen

In der Sprachhierarchie folgen auf die regulären Sprachen die als Typ-2-Grammatiken bezeichneten kontextfreien Sprachen. Diese sind in der Lage, symmetrische Einbettungen wie beispielsweise „Das Kind, das mit dem Ball spielte, hieß Peter." zu generieren. Formal würde dies mit $L=\{a^n bc^n | n \geq 0\}$ beschrieben, was bedeutet, dass zu jedem Element a (‚Das Kind') genau ein komplementäres Element c (‚hieß Peter') gehört. Die Elemente a und c gruppieren sich dabei symmetrisch um das Element b (‚das mit dem Ball spielte').

Diese Sprachen lassen sich mit Hilfe von als Stack oder Parser bezeichneten Kellerautomaten erkennen und decken bereits einen großen Teil der natürlichen Sprache ab. Viele Programmiersprachen sind typische Vertreter kontextfreier Sprachen und Parser zur syntaktischen Prüfung von Programmen im Rahmen der Umwandlung in Maschinencode sind Anwendungen dieser Automaten. Abbildung 6 zeigt am Beispiel eine Grammatik sowie deren Repräsentation als Syntaxbaum, dessen Knoten den Wortarten und Blätter den Worten entsprechen.

Kontextsensitive Sprachen oder Typ-1-Grammatiken sind in der Lage, Überkreuz-Relationen abzubilden, wie sie gelegentlich in der natürlichen Sprache vorkommen [Part+90, S.504]. Dazu muss der Akzeptor, in diesem Fall ein linear beschränkter Automat, auf die gesamte fragliche Zeichenfolge zugreifen können, um in Abhängigkeit umgebender Zeichen (dem Kontext) den entsprechenden Folgezustand einnehmen zu können.

Manche Programmiersprachen verwenden Deklarationsbereiche für Datentypen, anhand derer quasi kontextsensitiv entscheidbar ist, ob eine Variablenbelegung in diesem Kontext korrekt ist. Typ-0-Grammatiken haben nur theoretischen Wert, da ihr akzeptierender Automat einer Turingmaschine entspräche [Zaun99, S.21].

Natürliche Sprachen gehorchen ebenfalls einer Grammatik, weisen jedoch im unterschied zu formalen Sprachen unter Umständen sehr viele Unregelmäßigkeiten auf. Eine syntaktische Analyse auf Basis grammatischer Regeln ist umso schwieriger, je mehr Unregelmäßigkeiten eine Sprache aufweist. Auch die praktisch unbegrenzten Möglichkeiten zur Wortbildung aus dem Alphabet tragen dazu bei, dass sich natürliche Sprachen kaum durch formale Grammatiken angeben lassen [Lyons77, S.108].

Kontextsensitive Bestandteile natürlicher Sprache werden auf Grund des mit der Erkennung verbundenen Aufwands praktisch nicht in Betracht gezogen. Falls doch, wird der Kontext auf 1 bis 2 Worte im Umfeld begrenzt [Munr+03]. Man beschränkt sich beispielsweise in der syntaktischen Analyse auf kontextfreie Aspekte.

2.1.4.3 Computerlinguistik

Die vorliegende Arbeit widmet sich der Verarbeitung natürlicher Sprache durch Computer [Lepsk04]. Während sich der Forschungsbereich der Computerlinguistik diesem Thema aus sprachtheoretischer Sicht nähert, ist für das hier verfolgte Ziel eine technische Herangehensweise angezeigt. Um die Auswirkungen der Einbindung computerlinguistischer Methoden bewerten zu können, sind zumindest grundlegende Kenntnisse dazu hilfreich.

Die Computerlinguistik zielt auf die Erfassung syntaktischer und grammatischer Merkmale einer natürlichen Sprache, also der Gewinnung der Produktionen und Arten von Nonterminalen der Grammatik. Diese werden auch als Parts-of-Speech bezeichnet und dienen meist als externe Hinweise zur Erschließung von Texten (siehe S.81).

Neben den ab S.30 dargestellten statistischen, am Wort orientierten Verfahren zur Zurückführung der Flexions- auf die Grundformen, verwenden aktuelle Systeme wie beispielsweise Morphy [Lezi99] in starkem Maße Wörterbücher. Dort sind für jedes Wort in seiner Grundform die entsprechenden Flexionsregeln sowie Angaben zur Wort- und Silbentrennung bzw. Verfugung angegeben. An Hand dieser Einträge werden dann die zu einem Wort gehörige Grund- oder Stammformen sowie grammatische Merkmale wie z.B. die Wortart, ermittelt (siehe auch S. 30). Tritt eine Abfolge von Wortarten sehr häufig auf, so wird diese als gültige Produktion betrach-

tet und zur Erschließung weiterer, im Lexikon bisher nicht enthaltener Worte genutzt [KöAlt86].

Sowohl lexikonbasierte wie auch lexikonfreie Verfahren verfolgen das Ziel, Muster von Termfolgen zur Klassifikation weiterer Terme zu benutzen, wobei auf Grund der Komplexität der natürlichen Sprache ein hoher manueller Aufwand zur Verifikation der Produktionen sowie zur Erstellung und Pflege des Lexikons nötig ist [BeKö02]. Wird dieser Aufwand geleistet, sind mit derartigen Methoden zuverlässige syntaktische Analysen möglich. Es ist jedoch anzumerken, dass ein daraus resultierender Part-of-Speech-Tagger (POS-Tagger) nur für jeweils eine Sprache trainiert und anwendbar ist.

Die Verarbeitung natürlicher Sprache unter Verwendung vorgelagerter POS-Tagger wird in Abgrenzung zur Computerlinguistik als ***Natural Language Processing*** bezeichnet. Insbesondere im Rahmen der Informationsextraktion (siehe Abschnitt 2.5 ab Seite 79) werden syntaktische Informationen über die Wortart genutzt, wenn klar ist, dass die gesuchten Worte nur bestimmten grammatischen Klassen, also bestimmten Nonterminalen der Grammatik, angehören können [MaRa05].

Weiterhin lassen sich syntaktische Informationen vor- und nachfolgender Worte nutzen, um mit höherer Wahrscheinlichkeit entscheiden zu können, ob ein fragliches Wort der gesuchte Informationsträger ist. In dem Satz ‚Fed rose interest rates.' beispielsweise deutet das Substantiv nach ‚interest' auf die deutsche Entsprechung ‚Zins' an Stelle von ‚Interesse' hin [Fran02, S.14].

Viele existierende Systeme wie z.B. [Cali98], [CoSi99] zur Erkennung von Eigennamen in Texten nutzen daher die syntaktische Analyse durch POS-Tagger und verwenden die dort gewonnenen Informationen als externe Hinweise auf die Eigennamen. Dies bringt zwei Einschränkungen mit sich. Einerseits muss für jede zu bearbeitende Sprache ein entsprechender Tagger existieren und andererseits leidet die Vergleichbarkeit der Gesamtergebnisse bei Nutzung unterschiedlicher Systeme.

Das folgende Kapitel 2.2 widmet sich der Frage, mit welchen Methoden das traditionelle IR versucht, die dargestellten Herausforderungen durch die natürliche Sprache zu bewältigen. Dadurch sollen insbesondere Ursachen für Unzulänglichkeiten sowie geeignete, übertragbare Vorgehensweisen identifiziert werden.

2.2 Information Retrieval

Seitdem Menschen ihr Wissen aufschreiben, existiert auch die Aufgabe, das Aufgezeichnete wieder zu finden. Mit der zunehmenden Verbreitung von Literatur ergab sich die Notwendigkeit, dass an Stelle von Bibliothekaren die Information Suchenden selbst darauf zugreifen konnten. Es entstanden die ersten Verzeichnisse. Auf Grund der beschriebenen Schwierigkeiten natürlicher Sprache insbesondere auch bezüglich der Begriffsbildung war es dabei nahe liegend, sich am Wort, dem Indexwort, als der Bezeichnung eines Sachverhalts zu orientieren. Bezug genommen wurde jedoch nur auf diese konkrete Repräsentation, nicht den Begriff.

Die Erstellung des Index ist insbesondere deswegen problematisch, weil die zukünftigen Anfragen antizipiert werden müssen. Wichtige Aspekte der automatischen Indexierung sowie des Retrievals stellen die folgenden Abschnitte im Überblick dar.

2.2.1 Grundlegende Aspekte

Die zu Grunde liegende Aufgabe gliedert sich in zwei Teile: Zum einen die Erstellung der Indexe und zum anderen deren Nutzung, also der Zugriff auf die Dokumente beim Retrieval. Dabei lässt sich die Anfrageseite ohne Beschränkung der Allgemeinheit als Spezialfall der Indexierung auffassen, indem die Wortfolge der Anfrage als eigenständiges Dokument betrachtet wird. Die Beantwortung entspricht somit der Ermittlung der zur Anfrage ähnlichsten Dokumente.

2.2.1.1 Problembereiche

Das zentrale Problem des IR ist damit die Bestimmung der Ähnlichkeit zwischen Dokumenten. Legt man die Orientierung am Wort zu Grunde, so gilt es, Mengen von Worten zu vergleichen. Auf Grund der durch die Grammatik bedingten Flexionen müssen die zu betrachtenden Worte vorab auf eine einzige, alle anderen Formen repräsentierende Form abgebildet werden (vgl. Abschnitt 2.2.3.1).

Da wie dargestellt einzelne Worte nur mittelbar und oft in Abhängigkeit eines Anwendungsbereichs als gültiger Bezeichner eines Sachverhalts zu verstehen sind, müssen unter anderem Synonyme und Homonyme betrachtet werden. Dabei dürfen die zu vergleichenden Wortmengen nicht zu klein sein, damit beispielsweise

Polyseme auf Grund des gemeinsamen Auftretens mit anderen Worten genauer spezifiziert und Ambiguitäten aufgelöst werden können.

Daraus leitet sich die Frage ab, welche Mächtigkeit eine solche Menge von Worten aufweisen muss, um ein Dokument hinreichend genau zu beschreiben. Im Allgemeinen wird man möglichst viele Worte benutzen um damit eine hohe Zahl vergleichbarer Merkmale zu erhalten. Zur einfacheren Verarbeitung werden die Wortmengen in Form von Vektoren entsprechender Dimensionalität abgelegt.

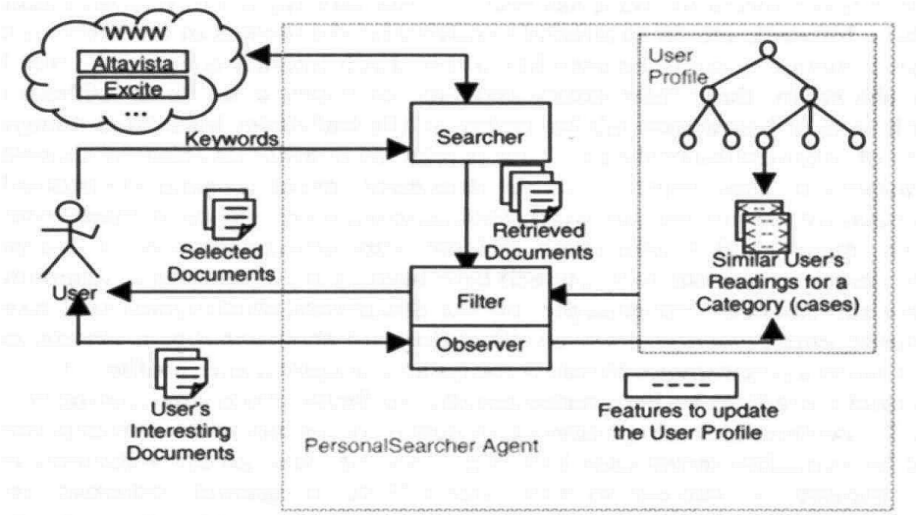
Gemäß der oben dargestellten Situation, zu einer Suchanfrage das dazu ähnlichste Dokument eines Korpus als Antwort zu identifizieren, wird der Aspekt der Dimensionalität besonders deutlich. Bei der Formulierung der Anfrage muss man sich also überlegen, welche Worte für den gesuchten und möglichst nur diesen Sachverhalt typisch sind [Kolo93, S.115]. Nur diese Worte gewährleisten eine zuverlässige Trennung zwischen relevanten und nicht relevanten Dokumenten.

Die Komplexität wird demnach auf die Seite der Anfrage verlagert. Um diese vor dem Benutzer zu verbergen wurden Verfahren entwickelt, die unter der Bezeichnung ***Query Expansion*** zusammengefasst werden. Dazu zählen ***Benutzerprofile***, ***Relevance Feedback*** und ***Collaborative IR***, die im Folgenden dargestellt werden.

2.2.1.2 Benutzerprofile

In einer Arbeit von Godoy und Amandi [GoAm00] wird ein elektronischer Agent eingesetzt um an Hand von Stichworten die Themen der vom Benutzer aufgerufenen Internetseiten zu identifizieren. Diese Informationen werden dann genutzt, um Ergebnisse von Suchanfragen auf interessierende Gebiete einzuschränken. Es wird also registriertes Interesse vergangener Anfragen genutzt, um das Retrieval auf ähnliche Fälle einzugrenzen [AaPl94]. Diese Vorgehensweise lässt sich leicht auf ein Szenario im Internet übertragen, indem die Anwender beispielsweise durch Cookies identifiziert und so gezielt beworben werden. Einen schematischen Überblick über das System gibt Abbildung 7.

Abbildung 7: Architektur des PersonalSearcher



Quelle: [GoAm00]

2.2.1.3 Relevance Feedback

Anfragen an Retrievalsysteme bestehen oft nur aus wenigen Worten, was zu einer großen Menge von Antwortdokumenten führt. Die Anfrage 'Relevance Feedback' bei Google™ beispielsweise bringt etwa 15,7 Millionen Treffer (Aufruf am 2005-10-21) unterschiedlichster Qualität im Sinne von Relevanz.

Die Beurteilung der Qualität der Ergebnisse durch den Nutzer wird als Relevance Feedback bezeichnet. Zu Aspekten der Wirkungsweise und empirisch belegten Modellen sei auf Harman [Harm92] verwiesen. Je nach Art der Erfassung der Einschätzung des Nutzers ist in explizites und implizites Vorgehen zu unterscheiden.

Beim expliziten Vorgehen beurteilt der Anwender die Menge der Ergebnisdokumente zu einer Anfrage q und teilt sie in die Menge der Relevanten R und der Irrelevanten I. Dabei bezeichnet q eine Menge von Suchtermen[9], die beispielsweise gleichmäßig gewichtet sind. Aus der Beurteilung der Relevanz aus Sicht des Nutzers ergibt sich die neue Anfrage q', die sich beispielsweise gemäß Formel 1 ermitteln lässt. Dabei bezeichnet v_d das Gewicht eines Elements der Indexmenge des Dokuments d und i bzw. r die Mächtigkeit der Mengen I bzw. R. Die Parameter α, β und γ sind anwendungsbezogen zu wählen und steuern, wie stark das Feedback die neue Anfrage beeinflusst.

9 Diese sind üblicherweise in Form eines Vektors organisiert, wobei jede Dimension einem Term entspricht. Details dazu werden in Abschnitt 2.3.1 ab Seite 54 dargestellt.

Das Gewicht v_d eines Indexterms gibt an, welche charakterisierende Bedeutung dieses Wort für den Text hat und ist Gegenstand des Abschnitts 2.2.3.2. Die Menge der neuen Suchterme *q'* sowie deren Gewichtung ergibt sich somit aus den Termen der alten Anfrage plus den gewichteten Indextermen der relevanten Dokumente minus der gewichteten Indexterme der irrelevanten Dokumente.

Formel 1: Bestimmung der neuen Anfragevektors beim Relevance Feedback

$$q' = \alpha q + \beta \left(\frac{1}{r} \sum_{d \in R} v_d \right) - \gamma \left(\frac{1}{i} \sum_{d \in I} v_d \right)$$

Quelle: [Ferb03, S.72]

Hat zum Beispiel der Term ‚halten' in der Menge der relevanten Dokumente die Gewichte 0.5, 0.2 und 0.7 und in der Menge der irrelevanten Dokumente die Gewichte 0.1, 0.9, 0.1 und 0.2, hat er in der neuen Anfrage das Gewicht 0.1, falls er in der ursprünglichen Anfrage nicht enthalten war und α, β und γ jeweils 1 sind. Je nach Implementierung sind negative Werte zu ignorieren oder werden als ‚darf nicht vorkommen' interpretiert.

Explizite Rückmeldungen zur Relevanz der Retrievalergebnisse werden zwar noch immer genutzt, bringen aber prinzipielle Schwierigkeiten mit sich. Zum einen sind nur wenige Anwender bereit, überhaupt eine Bewertung abzugeben, da dies mit zusätzlichem Aufwand verbunden ist. Zum anderen werden eher diejenigen Anwender Rückmeldung geben, die mit den Ergebnissen unzufrieden sind. Der Vorteil expliziter Rückmeldungen ist deren klare und direkt verwertbare Aussage.

Als implizit werden alle sonstigen, nicht expliziten Reaktionen des Nutzers auf die Ergebnisse des Retrieval bezeichnet. Insbesondere im Anwendungsbereich der Suchmaschinen im Internet bzw. Intranet kommen oft statistische Auswertungen zum Einsatz, da hier auf Grund der hohen Nutzerzahl eine hinreichende Signifikanz der Ergebnisse erhofft wird. Erfasste Größen sind dabei z.B. die Verweildauer auf einer Ergebnisseite oder die Anzahl der tatsächlich von der Ergebnisseite aus verfolgten Hyperlinks.

Pseudo Relevance Feedback als ein Beispiel zur rein maschinellen Bewertung von Relevanz geht davon aus, dass in den Ergebnislisten von Suchmaschinen die relevantesten Treffer die höchste Bewertung erzielen und nutzen daher die Ergebnisse mehrerer Suchmaschinen zur Verbesserung künftiger Anfragen.

Einen Mittelweg zwischen impliziten und expliziten Hinweisen auf Relevanz bilden fremde Referenzen. Dem zu Grunde liegt die Überlegung, dass Dokumente dann besonders relevant sind, wenn von vielen ähnlichen Dokumenten aus auf sie verwiesen wird. Gegen diese Vorgehensweise spricht die immer wieder zu beobachtende Methode, dass sich Autoren bzw. verschiedene Dokumente eines Autors damit gegenseitig Relevanz zusprechen [Bage05].

2.2.1.4 Collaborative Information Retrieval

Unter *collaborative Information Retrieval* versteht man das Zusammenfassen mehrerer Suchanfragen um die Antwortqualität zu erhöhen. Das Prinzip ist, dass Anfragen nach bestimmten Informationen von verschiedenen Personen ähnlich gestellt werden. War eine Antwort für eine Person nützlich, so wird unterstellt, dass bei ähnlicher Anfrage die gleiche Antwort auch von einer anderen Person als nützlich empfunden wird. Desto mehr Personen bei ähnlicher Anfrage diese Antwort als nützlich bewerten, desto wahrscheinlicher ist sie für folgende, ähnliche Anfragen nützlich.

Dieses Vorgehen lässt sich als Empfehlungssystem interpretieren [ShMa95], wie es auch im ***Case Based Reasoning*** genutzt wird [Berg+03, S.11] Damit einher geht jedoch unter Umständen eine Einebnung des Informationsangebots: diejenigen Dokumente werden als am relevantesten betrachtet, die für die größte Anzahl von Nutzern relevant waren. Im Umkehrschluss werden weniger oft genutzte Informationen als weniger relevant betrachtet, auch wenn dies im konkreten Fall falsch ist.

Bisher wurde sehr allgemein von der Qualität einer Antwort im Sinne ihrer Relevanz für eine Anfrage gesprochen. Diese wird zwar a posteriori subjektiv durch den Benutzer empfunden, muss aber operationalisiert werden um a priori abgeschätzt werden zu können. Dies wird im folgenden Abschnitt thematisiert.

2.2.2 Beurteilung der Qualität

Zur Beurteilung der Qualität eines IR-Verfahrens dienen zwei Kenngrößen. Dies ist einerseits die Vollständigkeit der auf eine Anfrage hin gefundenen Dokumente, genannt ***Recall***. Deren Relevanz für die Beantwortung der Suchanfrage wird als ***Precision*** bezeichnet. Diese Begriffe sind in Anlehnung an Ferber [Ferb03, S.86] wie folgt definiert:

Definition III: Relevanz, Relevanzrelation

Sei $D = \{d_1, ..., d_m\}$ eine gegebene Menge von Dokumenten und $Q = \{q_1, ..., q_n\}$ eine Menge von Anfragen. Sei weiter $R \subseteq DxQ$ die Teilmenge von DxQ, für die gilt: $(d,q) \in R \Leftrightarrow$ d ist relevant für q, dann wird R als Relevanzrelation auf DxQ bezeichnet.

Die Relation R, also die Beurteilung ob ein Dokument d für eine Anfrage q relevant ist, wird durch Befragung von Experten ermittelt. Praktisch bedeutet dies, dass eine Menge von Anfragen definiert werden. Damit entspricht die Relevanzrelation einer vorab definierten Erwartung bezüglich der Retrievalergebnisse. Um den Grad der Subjektivität zu verringern sind verschiedene Vorgehensweisen denkbar, die hier jedoch nicht erörtert werden sollen. Zu der ganz wesentlichen Frage, welchen Umfang die Mengen D und Q haben sollten und für wie viele Kombinationen daraus in der Relevanzrelation enthalten sein sollten, sind praktisch keine Angaben zu finden.

Definition IV: Precision und Recall

Sei $d \in D$ ein Dokument, $q \in Q$ eine Anfrage, R eine Relevanzrelation auf DxQ und D_q die Menge der in D zu q gefundenen Dokumente. Sei weiter $R_q = \{d \in D \mid (d,q) \in R\}$ die Menge der zur Anfrage relevanten Dokumente. Dann heißt

$$P(q,d) := \frac{|D_q \cap R_q|}{|D_q|}$$

Precision oder Genauigkeit und

2.2.2 Beurteilung der Qualität

$$R(q,d) := \frac{|D_q \cap R_q|}{|R_q|}$$

Recall oder Vollständigkeit der Antwort D_q auf die Anfrage q.

Werden auf eine Anfrage alle Dokumente eines Korpus zurückgegeben, so beinhaltet dies auch die tatsächlich Relevanten und es wird ein Recall von 1 erreicht. Umfangreiche Indexe fördern also hohe Recall-Werte. Precision hingegen ist maximal, falls ausschließlich relevante Dokumente geliefert werden. Hierfür sind gezielt gewählte, also aus Sicht der Klassifikation stark trennende Terme nötig, die entsprechend selten sind. Um diese schwer zu vereinbarenden Ziele zu verbinden, sind beide Größen zusammen zu fassen, was auch die Vergleichbarkeit unterschiedlicher Verfahren erleichtert. Dazu wird das als F-Measure bezeichnete harmonische Mittel beider Größen gemäß Formel 2 herangezogen, wobei P für Precision und R für Recall steht.

Formel 2: F-Measure

$$f = \frac{2 \bullet P \bullet R}{P + R}$$

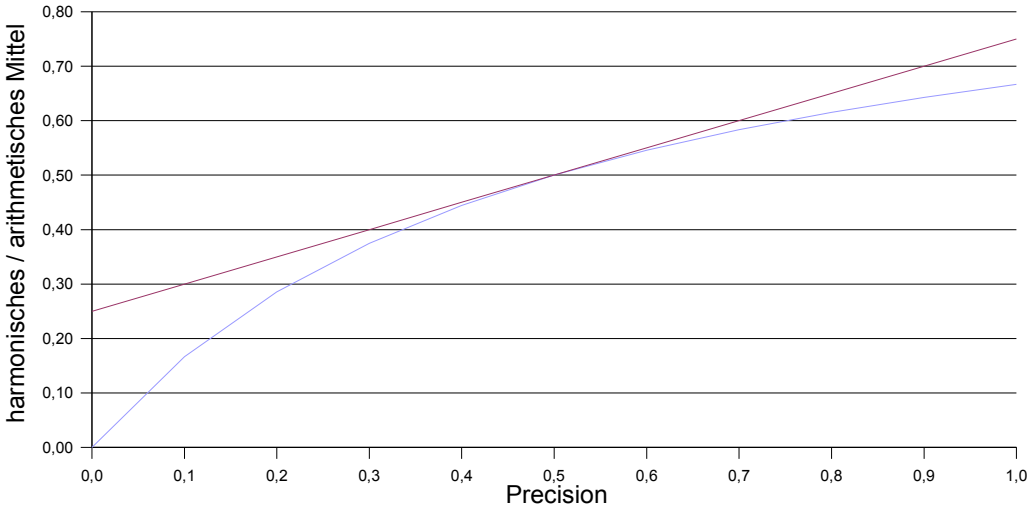
In dem in Abbildung 8 dargestellten Beispiel wurde Recall konstant auf 0,5 gehalten und die Entwicklung des Mittelwertes bei steigender Precision abgetragen. Im Gegensatz zum linearen Anstieg des arithmetischen Mittels wird also erreicht, dass ein geringer Wert eines Parameters weniger stark durch den Anstieg eines anderen Parameters kompensiert wird.

Auf Grund der empirischen Ermittlung der Relevanzrelation lassen sich Vergleiche unterschiedlicher IR-Methoden nur anhand von Referenzkorpora wie [Reut00] und den dafür definierten Anfragen anstellen. Aussagen über die unterschiedliche Qualität in anderen Szenarien sind nur begrenzt ableitbar [Strö04].

Antwortsysteme stellen eine spezielle Form von IR-Systemen dar, die auf eine Anfrage hin genau einen Textausschnitt von 50 oder 250 Zeichen Länge als Antwort liefern. Es wird also nur ein Ausschnitt eines für eine Anfrage relevanten Dokuments zurückgegeben, obwohl unter Umständen mehrere Dokumente die ge-

wünschte Information beinhalten. Die maschinelle Bewertung der Richtigkeit einer Antwort gestaltet sich daher schwierig.

Abbildung 8: Harmonisches Mittel im Beispiel



Quelle: eigene Erstellung

Auch hier wird versucht, mit Hilfe von Referenzkorpora, Fragen und zugehörigen Antworten die Qualität der Retrievals zu messen. Allerdings gilt noch immer

> *"Creating the true equivalent of a standard retrieval test collection is an open problem."*
>
> *[TREC04]*

Zusammenfassend ist festzustellen, dass die dargestellten Methoden zur Beurteilung der Qualität zum Vergleich verschiedener Verfahren im Information Retrieval nur dann aussagekräftige Ergebnisse bereitstellen, wenn das Einsatzszenario konstant gehalten wird.

Für die Evaluation des hier zu entwickelnden Verfahrens bedeutet dies, dass zu untersuchen ist, wie sich die Ergebnisse des Verfahrens bei Änderung einzelner Parameter des Szenarios entwickeln. Diese Parameter sind beispielsweise Sprache und Umfang des betrachteten Korpus.

2.2.3 Indexierung

Natürliche Sprache bildet durch die Vielzahl möglicher Ausdrucksformen und Bezeichnungen ein sehr komplexes System mit vielen kontextabhängigen Bezügen. Eine vollständige syntaktische Analyse wie in Kapitel 2.1.4.2 dargestellt ist daher sehr aufwändig und fehleranfällig. Im Gegensatz dazu ist eine Volltextsuche sehr einfach. Diese ist aber nur dann geeignet, wenn lediglich eine geringe Zahl von Dokumenten in Frage kommt, so dass ein schneller Zugriff gewährleistet ist. Weiterhin muss der Information Suchende mit den in den Texten verwendeten Worten vertraut sein, da nur exakte Übereinstimmungen zwischen zwei Zeichenfolgen gefunden werden.

Um aus beiden Extremen einen geeigneten Kompromiss zu bilden muss einerseits die Komplexität der Sprache reduziert und müssen andererseits Bezeichnungen generalisiert werden [Fugm92]. Dies geschieht durch Abbildung auf eine Menge von Indextermen. Die dazu verwendeten Verfahren werden im folgenden dargestellt.

2.2.3.1 Wörterbuch- und Regelverfahren

Als Wörterbuch wird in diesem Zusammenhang eine a priori festgelegte Liste von Worten bzw. Wortbestandteilen verstanden. Neben der Verwendung als Stoppwortliste zur Volumenverringerung dienen sie wie Regelverfahren der Kompositazerlegung und der Reduktion auf Grund- bzw. Stammformen.

Die Stammform eines Wortes entsteht durch Abtrennen von Vor- und Nachsilben, während die Grundform die unflektierte Form des Wortes ist. So ist ‚grub' die Stammform von ‚vergrub' und ‚graben' dazu die Grundform. Beide Formen der Reduktion bewirken also eine Generalisierung. Kontextsensitivität kann zu einem gewissen Grad durch Regeln oder die Betrachtung von Wortgruppen erreicht werden. Oft werden daher Kombinationen aus Regelsystemen und Wortlisten genutzt [Hahn+01].

Gemeinsamer Nachteil dieser Verfahren ist deren geringe Flexibilität und die große Fehleranfälligkeit auf Grund des oft hohen Erstellungsaufwands. Folgendes Beispiel von Hauser verdeutlicht dies:

> *„Die 40.000 am häufigsten verwendeten Worte des Deutschen gliedern sich in 23.000 Substantive, 6.000 Verben sowie 11.000 Adverbien und Adjektive. Substantive kennen 4 grammatische Fälle, Verben 24 und Adjektive / Adverbien 18 Flexionsformen. Diesen 40.000 Worten stehen also etwa 434.000 mögliche Formen gegenüber!“*[10]
>
> *[Haus99, S.245]*

Vor allem für sehr regelmäßige Sprachen lassen sich jedoch mit überschaubarem Aufwand Regelsysteme aufbauen. Für das Englische beispielsweise hat Kuhlen ein solches System zur Bildung von Grundformen vorgestellt, was sehr gute Ergebnisse liefert [Kuhl77, S.11].

Vor allem aus sprachtheoretischer Sicht wird immer wieder auf die geringe Eignung von Wortlisten zur Beschreibung von Inhalten hingewiesen [Pole78, S.164]. Wegen der einfachen Handhabbarkeit finden sie jedoch sehr häufig Verwendung. Ihre wichtigsten Anwendungen sind:

- Stoppwortlisten: Terme ohne Bedeutung für das Verständnis des Textes werden entfernt. Dies sind insbesondere Artikel, Pronomen und Füllworte, aber auch solche Worte, die innerhalb einer Domäne wenig bezeichnend sind.
- Trunkierung: Bei dieser Form der Stammformbildung werden Vor- und / oder Nachsilben eines Wortes entfernt. Bei Linkstrunkierung werden Vorsilben, bei Rechtstrunkierung die Suffixe an Hand einer vorgegebenen Liste abgetrennt.
- Kompositazerlegung: Aufspaltung zusammengesetzter Worte in die Einzelworte;
- Thesauri: Sie bilden Verbindungen zwischen Worten ab. Da ein Thesaurus trotz der prinzipiellen Listenform auch weiterreichende Zusammenhänge beinhaltet, soll darauf genauer eingegangen werden.

In einem Thesaurus werden die für ein Sachgebiet relevanten Terme und deren Verbindung untereinander sowie ergänzende Informationen durch Merkmal-Wert-Paare erfasst. Gemäß DIN 1463-1 sind folgende Spezifikationen vorgesehen:
BF – Benutze für Synonym; BS – Benutze Synonym; OB – Oberbegriff; UB – Un-

10 Komposita mehrerer Worte sind dabei noch nicht berücksichtigt.

terbegriff; VB – Verwandter Begriff (siehe Abbildung 9). Oft werden auch Antonyme sowie eine kurze Beschreibung des betroffenen Begriffs aufgenommen.

Abbildung 9: Beispiel für Thesauruseintrag 'Dokumentationssprache'

Merkmal	Wert
BF	Indexierungssprache
OB	Künstliche Sprache
UB	Klassifikationssystem
UB	Thesaurus

Quelle: [INFO05]

Der Unterschied zwischen BF und BS liegt dabei in der Unterscheidung in Deskriptor, also Indexterm, und Nichtdeskriptor. Im Beispiel ist der Deskriptor ‚Dokumentationssprache', der für das Synonym ‚Indexierungssprache' zu benutzen ist. Typischerweise ist der Deskriptor die exakte Bezeichnung und das Synonym eine Verallgemeinerung.

Thesauren definieren damit ein kontrolliertes Vokabular und stellen Beziehungen zwischen dessen Termen her. Sie bieten so einen Ansatz zur Strukturierung von Inhalten um gezielt auf Informationen zugreifen zu können.

Da die manuelle Erstellung von Wortlisten, also die Definition der zu verwendenden Indexterme sehr aufwändig und damit teuer ist, versucht man mit Hilfe statistischer Verfahren diesen Vorgang zu automatisieren. Die dazu verwendeten Verfahren sind Gegenstand der folgenden Abschnitte.

2.2.3.2 Basis statistischer Verfahren

Das Ziel statistischer Verfahren ist es, diejenigen Worte zu identifizieren, die den Inhalt des Dokuments am besten beschreiben. Ihren Ausgangspunkt haben diese Verfahren in den Überlegungen von Georg Zipf (1902-1950). Dieser stellte fest, dass sich die Häufigkeit h und der Häufigkeitsrang i der einzelnen Worte beliebiger Textsammlungen gemäß $h = i^{-a}$ verhält, wobei a etwas größer als 1 ist [Zipf32, S.8-25]. Häufigkeit und Häufigkeitsrang verhalten sich also indirekt proportional zueinander. Dieses Phänomen wird als Zipf-Verteilung bezeichnet.

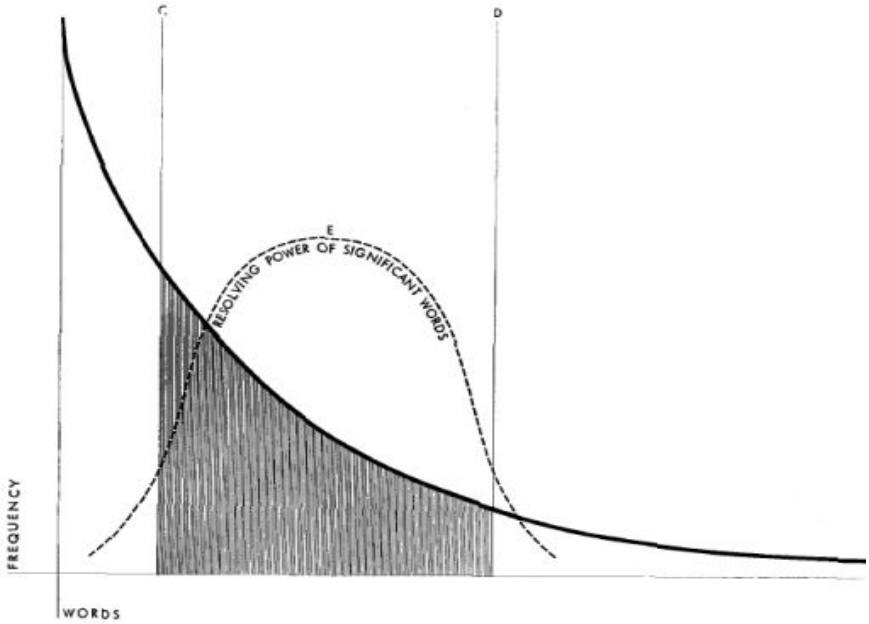
Nutzbar ist dieser Zusammenhang, wenn man den häufigsten sowie den seltensten Worten unterstellt, nicht zur Beschreibung des Dokuments beizutragen. Allerdings verhindert die Angabe konkreter Werte für h eine Verallgemeinerung auf andere Korpora. Mandelbrot [Mande53] untersuchte diesen Zusammenhang genauer und erweiterte ihn im Sinne einer Vorkommenswahrscheinlichkeit P(i) gemäß Formel 3 mit den korpusbezogen empirisch zu ermittelnden Konstanten c und b.

Formel 3: Zipf-Mandelbrot-Gesetz

$$P(i) = \frac{c}{(i+b)^a}$$

Dadurch sind Aussagen möglich wie ‚Worte mit einer Auftretenswahrscheinlichkeit zwischen 0,3 und 0,8 sind zur Beschreibung eines Dokuments geeignet' [Luhn58]. In Abbildung 10 ist dies dargestellt. Worte höherer Wahrscheinlichkeit sind demnach als Stoppworte zu betrachten, während seltenere Worte als zu spezifisch und damit ebenfalls wenig aussagekräftig bewertet werden.

Abbildung 10: Relevanter Bereich der Zipf-Verteilung



Quelle: [Luhn58]

Diese Betrachtung von Wahrscheinlichkeiten legt nahe, dass die zu Grunde liegenden Textsammlungen entweder einen gewissen Mindestumfang aufweisen oder die enthaltenen Worte zumindest vorab auf Grund- oder Stammformen reduziert worden sind. Nur so ist zu gewährleisten, dass signifikante Unterschiede der

Häufigkeiten auftreten. Außerdem nimmt diese Herangehensweise keine Rücksicht auf Einzeldokumente. Worte, die bezogen auf ein einzelnes Dokument dem Häufigkeitskriterium genügen, aber für den Gesamtkorpus keine Rolle spielen werden nicht berücksichtigt.

Eine auf Häufigkeitsuntersuchungen aufbauende Vorgehensweise, die diesen Dokumentbezug beinhaltet ist die Ermittlung der als TF-IDF (Term Frequency – Inverse Document Frequency, Formel 4) bezeichneten Kenngröße. Sie verknüpft die Auftretenshäufigkeit eines Terms in einem Dokument mit der relativen Anzahl von Dokumenten, die diesen Term enthalten.

Formel 4: Ermittlung der gewichteten Kenngröße TF-IDF

$$tfidf_{d,t} = c + (1-c) * \frac{f_{d,t}}{\max\{f_{d,t_i} \mid t_i \in T\}} * \log(f_d / f_t)$$

f_d - Anzahl Dokumente; f_t - Anzahl Dokumente mit Term t; $f_{d,t}$ - Vorkommen von t in d

TF-IDF ermittelt somit Terme, die in wenigen Dokumenten des Korpus auftreten, innerhalb dieser Dokumente jedoch häufig sind. Damit nicht lange Dokumente bevorzugt werden, findet unter Umständen eine Gewichtung der Termhäufigkeiten mit einem Faktor $c \in [0;1]$ sowie eine Normalisierung auf den häufigsten Term des Dokuments oder auf die Gesamtzahl der Worte eines Dokuments statt. Als relevant werden dann alle Worte betrachtet, deren Werte für TF-IDF größer als ein empirisch zu ermittelnder Grenzwert sind.

Die dargestellten statistischen Verfahren werden dazu benutzt, den Inhalt von Dokumenten mit Hilfe von Wortmengen zu beschreiben. Diese Wortmengen werden im Information Retrieval dazu benutzt, Themen zu charakterisieren und somit Dokumente zu klassifizieren [Net+00]. Ein Thema wird dabei durch eine Menge gemeinsam auftretender Worte beschrieben. Enthält ein Dokument hinreichend viele für ein Thema typische Worte, wird dieses entsprechend thematisch eingeordnet. Damit kann ein Korpus nach Themen strukturiert und somit der für eine Anfrage relevante Suchraum verkleinert werden.

2.2.3.3 Kollokationen

Kollokationen bezeichnen das statistisch signifikante gleichzeitige Auftreten von Termen. Diese können in beliebiger Entfernung, beispielsweise innerhalb eines Satzes oder eines Dokuments vorkommen. Man unterstellt diesen Termen die Bildung einer sinnvollen Einheit oder zumindest einen semantischen Zusammenhang [Heye+01, Biem+03]. In unmittelbarer Nachbarschaft wie bei ‚New York' stellen sie potenzielle Elemente der Indexmenge dar. Entsprechend der Anzahl N unmittelbar aufeinander folgender Worte wird auch von N-Grams gesprochen, ‚New York' ist demnach ein 2-Gram.

In der Computerlinguistik wird oft mit Hilfe von Kollokationsanalysen bekannter Zusammenhänge auf die Wortklasse eines konkreten Wortes geschlossen [Fran02]. Meist wird dabei eine statistisch ermittelte Wahrscheinlichkeit für die Zuordnung zu Grunde gelegt, um Unsicherheiten, vor allem auf Grund von Ambiguitäten, Rechnung zu tragen.

Im Information Retrieval werden Kollokationen unter anderem zur Bewertung der Ähnlichkeit von Dokumenten verwendet. Tauchen in einer Textsammlung beispielsweise sehr häufig die Worte ‚Zahnarzt' und ‚Karies' innerhalb eines Absatzes auf, so unterstellt man einen inhaltlichen Zusammenhang. Darauf aufbauend werden Dokumente als ähnlich betrachtet, sofern sie ähnliche Kollokationen aufweisen. Zur Begrenzung des Aufwandes beschränkt man sich auf bestimmte, vorab als relevant für den gegebenen Diskursbereich definierte Terme (siehe oben).

Zur Beurteilung der Stärke des Zusammenhangs zwischen zwei Termen i und j wird untersucht, wie häufig deren gemeinsames Auftreten h(i,j) im Verhältnis zum Auftreten jedes einzelnen Wortes h(i) bzw. h(j) ist.

Formel 5: DICE-Maß

$$DICE(i,j) = \frac{2 * h(i,j)}{h(i) + h(j)}$$

In der Praxis haben sich vor allem das dem harmonischen Mittel (vgl. Formel 2) ähnliche DICE-Maß nach Formel 5 sowie das in Formel 6 dargestellte Jaccard-Maß etabliert.

2.2.3 Indexierung

Formel 6: Jaccard-Maß

$$Jaccard(i,j) = \frac{h(i,j)}{h(i)+h(j)-h(i,j)}$$

Auch in der Kollokationsanalyse findet sich die Orientierung an festen, als relevant angenommenen Einzelworten und nicht am durch dieses Wort repräsentierten Begriff wieder. Das statistisch signifikante gemeinsame Auftreten der unterschiedlichen Worte wird jedoch zumindest als eine Möglichkeit zur Beschreibung des in einem Dokument behandelten Themas bzw. der Begriffe im Sinne des extensionalen Begriffsverständnisses benutzt [Ferb03, S.225].

Eine derartige Vorgehensweise haben Quasthoff und Wolff [QuWo02] zur Erstellung von ***Topic Maps*** verwendet. Diese Themenkarten werden zur grafischen Darstellung von Zusammenhänge mit Hilfe hyperbolischer Bäume genutzt. Sie basieren auf durch Kollokationsanalysen gewonnene Informationen über Zusammenhänge relevanter Terme eines Bereichs.

Es ist offensichtlich, dass die so postulierten Zusammenhänge von der zu Grunde liegenden Textkollektion abhängen und damit nur bedingt übertragbar sind. Komplexere Verfahren betrachten darüber hinaus den verwendungshistorischen und/oder syntaktischen Zusammenhang in dem ein Term auftritt [Gref92]. Einen Überblick bietet [Ferb03, S.223ff].

Die Betrachtung von Häufigkeiten hat zur Folge, dass stets auch Fehler auftreten. Damit eignet sich diese Herangehensweise vor allem für umfangreiche Textsammlungen, da nur so ein hinreichend großer Stichprobenumfang gewährleistet ist. Dieser ist Voraussetzung dafür, dass einzelne Fehler an Gewicht verlieren. Der durch die unnötigen Indexterme entstehende rechnerische Aufwand relativiert sich durch die Verfügbarkeit von Computern, so dass diese Vorgehensweise sehr weit verbreitet ist.

Auf Grund ihrer einfachen Nachvollziehbarkeit und damit Evaluierbarkeit finden statistische Ansätze in sehr vielen Bereichen der Textanalyse Verwendung. Insbesondere sei hier die quantitative Linguistik als ein Teilgebiet der Computerlinguistik erwähnt [Haus99, MaSch00].

Eigene Untersuchungen zur Analyse eines deutschsprachigen Korpus mit statistischen Kenngrößen finden sich am Ende dieses Kapitels. Trotz der dabei getroffenen Vereinfachungen wird die Bedeutung des Stichprobenumfangs deutlich: Um mit Hilfe eines statistisch konstruierten Index gezielte Zugriffe zu erreichen, muss dieser relativ umfangreich sein. Außerdem sind Wörterbücher oder grammatische Regeln notwendig, die die Menge der Flexionsformen reduzieren. Die Mehrzahl der Worte tritt sonst nur ein oder zwei Mal auf, so dass diese statistisch irrelevant sind. Allerdings sind die meisten Nutzer mit dieser Vorgehensweise beim Retrieval vertraut und wählen selbständig alternative Suchworte, falls die Anfrage unbefriedigend verlief.

Die Interpretation der Auftretenshäufigkeit $h(x_i)$ als Wahrscheinlichkeiten ist kritisch zu sehen. Zwar ist die Bestimmung der Auftretenswahrscheinlichkeit P eines Wortes x_i unproblematisch als

$$P(x_i) = \frac{h(x_i)}{\sum_{k=1}^{n} h(x_k)}$$

möglich. Fraglich ist jedoch, wie diese Wahrscheinlichkeiten zu verknüpfen sind, was letztlich mit der Frage einher geht, ob es sich um stochastisch abhängige oder unabhängige Ereignisse handelt. Letzteres würde bedeuten, dass Kommunikation nichts als eine zufällige Aneinanderreihung von Worten ist. Um einen Sachverhalt auszudrücken wählt man aber nicht zufällig ein Wort aus dem Wortschatz, sondern berücksichtigt unter anderem grammatische Regeln sowie die im Verlauf bereits genutzten Worte. Die damit einhergehende stochastische Abhängigkeit lässt sich aber weder widerlegen noch beweisen, da sich stets Fälle für und wider dieser Hypothese finden lassen [Coop91].

Die Wahrscheinlichkeitstheorie bietet jedoch viele nützliche Hilfsmittel, so dass die Interpretation als Wahrscheinlichkeit zweckmäßig ist [Coop91, BoTi+03, S.296f]. Außerdem treten bei der Erschließung von Informationen, insbesondere bei der Klassifikation von Worten hinsichtlich ihrer Relevanz, Unsicherheiten auf, die sich durch die Verwendung von Wahrscheinlichkeiten abbilden lassen.

Unter der Bezeichnung ***Machine Learning*** werden in diesem Zusammenhang Verfahren zusammengefasst, die Klassifikations- bzw. allgemein Entscheidungsregeln aus Wahrscheinlichkeitsanalysen erstellen [MaSch00]. Damit gehen die Möglichkeiten der im Folgenden dargestellten probabilistischen Modelle über die reine Indexbildung hinaus [Rabi89]. Es ist jedoch darauf hinzuweisen, dass diese rein statistisch orientierte Auffassung des ***Machine Learning*** dem derzeit vorherrschenden Paradigma der statistischen Sprachverarbeitung geschuldet ist. Grundsätzlich fallen in dieses Gebiet alle Verfahren die darauf abzielen, mit Computern selbsttätige Erschließung von Informationen zu erreichen [Mitch97].

2.2.4 Probabilistische Modelle

In der Literatur wird oft zwischen probabilistischen und regelbasierten Modellen unterschieden [Klat02], wobei bei näherer Betrachtung durchaus Gemeinsamkeiten deutlich werden. Sehr oft werden Regeln auf Grund von häufig beobachteten Zusammenhängen postuliert, die jedoch nur mit einer bestimmten Wahrscheinlichkeit gültig sind.

2.2.4.1 Assoziationsregeln

Ihren Ursprung haben Assoziationsregeln (***Association Rules***) im Data Mining und hier insbesondere in der Warenkorbanalyse. Entsprechend wird in der Definition von einer Transaktion gesprochen, wenn eine Auswahl von Elementen gemeint ist. Dabei interessiert lediglich das Vorhandensein eines Elements, nicht aber dessen Kardinalität.

Assoziationsregeln dienen der Erkennung von signifikanten Zusammenhängen zwischen Elementen einer Menge. Bezogen auf den Bereich des Information Retrieval steht im Gegensatz zu Kollokationen hier die Betrachtung mehrerer Worte im Mittelpunkt.

Nach Aggarwal und Yu [AgYu98] liegen folgende Definitionen zu Grunde:

Definition V: Assoziationsregel

> *Sei $I=\{i_1, i_2, \ldots i_n\}$ eine Menge binärer Elemente. Seien $X \subseteq I$ sowie $Y \subseteq I$ disjunkte, als Transaktion bezeichnete Teilmengen von I. Dann wird ein Ausdruck der Form $X \Rightarrow Y$ als Assoziationsregel bezeichnet.*

Die Bezeichnung ‚X impliziert Y' wird vermieden, da im Gegensatz zur logischen Implikation Assoziationsregeln den Grad der Gültigkeit dieses Zusammenhangs betrachten. Einen Ansatz zur Bewertung der Gültigkeit gibt Definition VI.

Definition VI: Support, Confidence

*Der Anteil aller Transaktionen, die sowohl X als auch Y beinhalten, wird als **support**(X⇒Y) bezeichnet.*

Das Verhältnis

$$\frac{|X \cup Y|}{|X|}$$

*wird als **confidence**(X⇒Y) bezeichnet.*

Die minimal geforderten Werte für ***support*** und ***confidence***, oberhalb derer eine Assoziationsregel sinnvoll als gültig angenommen wird, sind anwendungsbezogen empirisch zu ermitteln. Im Gegensatz zur üblichen Bedeutung steht | . | hier nicht für die Mächtigkeit der Menge sondern für die Häufigkeit, mit der diese Menge vorkommt.

Das folgende Beispiel verdeutlicht dies sowie die Anwendung von Assoziationsregeln: In einem Korpus aus 1000 Dokumenten treten die Worte `Welt`, `Erde` und `Globus` in 500 Dokumenten gemeinsam auf, wobei in 400 dieser Dokumente gleichzeitig noch das Wort `Umwelt` auftritt. Damit ist X={`Welt`, `Erde`, `Globus`} mit | X | = 500 und Y={`Umwelt`} mit | X∪Y | = 400. Daraus lässt sich die Assoziationsregel `Welt`, `Erde`, `Globus => Umwelt` mit ***support*** von 0,4 und ***confidence*** von 0,8 ableiten. Damit erscheint es zweckmäßig, eine Retrievalanfrage mit den Termen `Welt`, `Erde` und `Globus` um den Term `Umwelt` zu ergänzen um so bessere Ergebnisse zu erzielen.

Zu lesen ist dieses Ergebnis wie folgt: „Die Wahrscheinlichkeit, dass in einem Dokument diese vier Terme gemeinsam auftreten beträgt 40%. Treten die Terme aus dem linken Teil der Regel in einem Dokument auf, so wird mit 80%-iger Wahrscheinlichkeit auch der rechte Teil der Regel auftreten." Anders ausgedrückt: unter der Bedingung, dass der linke Teil der Regel eingetreten ist, tritt der rechte Teil mit der durch Wahrscheinlichkeit von ***confidence*** auf. Es handelt sich also um die Er-

mittlung von Wahrscheinlichkeiten, wobei ***confidence*** der bedingten Wahrscheinlichkeit nach Bayes gemäß Formel 7 entspricht. Dieser Aspekt wird im folgenden Abschnitt nochmals aufgegriffen.

Formel 7: Bedingte Wahrscheinlichkeit nach Bayes

$$P(Y \mid X) = \frac{P(X \cap Y)}{P(X)}$$

Das angeführte Beispiel macht auch deutlich, dass Assoziationsregeln mit hohen Werten für ***confidence*** im Sinne des extensionalen Verständnisses zur Beschreibung von Begriffen genutzt werden können [Sing+99]. Aufbauend auf das Beispiel könnten Terme als weitere Instanzen des Begriffs `WELT` betrachtet werden, deren Assoziationsregeln ***confidence***-Werte von über 0,8 aufweisen. Mit Hilfe abgestufter Anforderungen an ***support*** und ***confidence*** lassen sich auch unscharfe Aussagen der Art ‚ziemlich relevant' oder ‚weitgehend ähnlich' modellieren [Wolf01], so dass verschiedene Detailierungsgrade bzw. Abstraktionsstufen eingeschlossen werden. Bezogen auf das IR entspricht dies Aussagen wie ‚Diese Worte sind sehr relevant für dieses Thema'.

Es ist anzumerken, dass die Ermittlung dieser Regeln sehr aufwändig ist, da für jede Kombination von Indextermen zu untersuchen ist, welche sonstigen Terme sich damit erschließen lassen. Werden nur die Terme konkreter Anfragen benutzt, so müssen diese mit allen anderen Indextermen in Verbindung gebracht werden.

Problematisch ist weiterhin, dass Regeln mit hohem ***support*** bekannt und somit uninteressant sind. Regeln mit geringem ***support*** kommen sehr häufig vor und sind daher oft irrelevant.

Ein weiterer, häufig genutzter Vertreter probabilistischer Modelle ist die im folgenden vorgestellte Entropiebetrachtung.

2.2.4.2 Informationstheoretischer Ansatz

Aus der Informationstheorie nach Shannon [Shan48] stammt ein Ansatz, der auf der Wahrscheinlichkeit des Auftretens eines Signals beruht und als Entropie bezeichnet wird. Information wird bei diesem Ansatz als beseitigte Unsicherheit be-

trachtet: desto geringer die Entropie, desto weniger Ungewissheit besteht. Entsprechend gilt für eine Signalquelle hoher Entropie, dass die betrachteten Signale stochastisch unabhängig sind und somit jedes neue Signal einen Informationsgewinn darstellt [Maur03, S.10f].

Der nützliche Informationsgehalt einer Signalquelle X mit den möglichen Signalen x_i und deren Auftretenswahrscheinlichkeit $p(x_i)$ mit $\sum_i p(x_i) = 1$ berechnet sich gemäß Formel 8. Oft wird der gesamte Term auch negiert und p(x) an Stelle von $p(x)^{-1}$ notiert. Bei Verwendung des dualen Logarithmus erhält man die in der Computertechnik übliche Einheit ‚Bit' [Maur03, S.7].

Formel 8: Entropie einer Informationsquelle

$$H(X) = \sum_{x \in X} p(x) \cdot \log\left(\frac{1}{p(x)}\right)$$

Damit entspricht eine Signalquelle einer Zufallsvariablen. Verbindet man zwei oder mehr Signalquellen X, ..., Z, so lässt sich diese Verbindung als einzelne, vektorielle Zufallsvariable X...Z mit der Verbundentropie H(X...Z) auffassen, die die Werte des kartesischen Produkts der einzelnen Signalquellen bzw. Zufallsvariablen annimmt [Maur93, Ebel+98, S.89]. Für die Untersuchung des gemeinsamen Auftretens von Signalen ist die bedingte Entropie und darauf aufbauend die gegenseitige Information zweier Signalquellen entsprechend Definition VII interessant [DaUt04]. Diese Größen ergeben sich aus den Formeln 7 und 8.

Definition VII: bedingte Entropie, gegenseitige Information

Die bedingte Entropie einer Signalquelle X unter Maßgabe des Signals der Signalquelle Y ist gegeben durch

$$H(X \mid Y) = H(XY) - H(Y)$$

und die gegenseitige Information, die X über Y gibt durch

$$I(X;Y) = H(X) + H(Y) - H(XY)\ .$$

Dabei entspricht I(X;Y) der Reduktion der Unsicherheit über X, wenn man Y erfährt.

Im Umfeld des Information Retrieval werden unter Signalen typischerweise Worte eines Dokuments verstanden. Der zentrale Punkt dieser Herangehensweise ist die Fragestellung, was sich über kommende Signale sagen lässt, wenn bereits einige Signale eingetroffen sind. Anders ausgedrückt: Wie viel Information gewinnt man durch das Eintreffen von Wort Y, falls zuvor das Wort X auftrat [HeHo98]. Dabei ist wie bei den dargestellten Assoziationsregeln im Allgemeinen X eine Menge von Worten. Bei der Anwendung im IR werden diese Kenngrößen zur Bewertung der (semantischen) Zusammengehörigkeit von Worten eingesetzt [Lav+04].

2.2.4.3 Bewertung probabilistischer Verfahren

Assoziationsregeln haben gegenüber anderen probabilistischen Verfahren den Vorteil, dass sie unmittelbar verständlich sind. Daher fällt es einem sachkundigen Entwickler leicht, sinnvolle und interessante Regeln zu identifizieren, die dann im Rahmen eines ***Query Expansion*** (siehe S. 23) genutzt werden können. Notwendig ist dazu jedoch eine Einschränkung der Anzahl zu betrachtender Regeln durch Vorgabe hinreichend hoher Schwellwerte für ***support*** und ***confidence***.

Interessant sind Regeln im Sinne der extensionalen Begriffsbildung insbesondere zur Erschließung in Frage kommender Wortmengen (vgl. S. 34). Hierzu kann von eher kleinen Mengen bekannter Bezeichnungen ausgehend untersucht werden, welche weiteren Terme typischerweise in Verbindung mit den Gegebenen auftreten.

Eine Übertragbarkeit einmal gefundener Regeln auf andere Diskursbereiche ist kaum möglich, da sie vom verwendeten (Fach-)Vokabular abhängen. Dieser jeweils zu erbringende manuelle Aufwand spricht gegen eine Nutzung im Rahmen der hier angestrebten Vorgehensweise.

Ein wesentlicher Nachteil aller auf Wahrscheinlichkeiten basierenden Verfahren liegt im notwendigen großen Umfang der zu Grunde liegenden Textsammlungen, die für eine zuverlässige Überführung der Statistik zu stochastischen Größen nötig ist. Daher wird oft mit Hilfe vorgeschalteter Verfahren zur Rückführung der Worte auf Stammformen die morphologische Vielfalt stark reduziert. Dieser Punkt steht ebenfalls zu den hier gesetzten Rahmenbedingungen im Widerspruch, da dies wie in Abschnitt 2.1.4.3 dargestellt nicht sprachneutral geschehen kann.

Dieser Nachteil spricht auch gegen die Verwendung des Entropieansatzes. Damit lassen sich zwar sehr gut typische Abfolgen von Worten bzw. Begriffen erkennen. Allerdings ist dazu ein hoher manueller Trainingsaufwand nötig, um tatsächlich von einer Wahrscheinlichkeit sprechen zu können [Lav+04].

Allen Verfahren gemeinsam ist, dass empirisch zu ermittelnde Schwellwerte notwendig sind, die die Stärke der Verbindungen zwischen Worten oder die Eignung eines Wortes als Indexterm abschätzen lassen. Je nach tatsächlicher Verteilung der ermittelten Werte erscheint hier eine relative oder kardinale Angabe geeignet. Beispielsweise ‚Die 10 Worte mit dem höchsten TF-IDF-Wert sind Indexterme' oder ‚Alle Regeln mit mindestens 95% des maximal erreichten ***confidence***-Wertes werden ausgewählt'.

Neben der Sicherung der Aussagekraft als Entscheidungskriterium für Relevanz im Sinne von Eignung haben diese Schwellwerte den Effekt, Teile des Suchraums abzutrennen. Dies kann zu einer Reduzierung des Berechnungsaufwands und somit höherer Effizienz der weiteren Verarbeitung führen.

2.2.5 Statistische Korpusdaten am Beispiel

Um die dargestellten statistischen bzw. probabilistischen Verfahren beurteilen zu können, werden im folgenden einige Kennzahlen am Beispiel ermittelt. Dabei wird entsprechend der Zielvorgaben dieser Arbeit ***keine*** Vorverarbeitung durch syntaktische Analysen durchgeführt.

2.2.5.1 Grundlegende Kenngrößen

Zur Darstellung des Häufigkeitsranges nach Zipf wurde das Jahresarchiv 2003 des Newstickers von heise-online, bezeichnet als 'K2003', herangezogen. Details zu diesem sowie anderen im weiteren Verlauf benutzten Korpora zeigt Tabelle 1.

Tabelle 1: Kenngrößen der Korpora

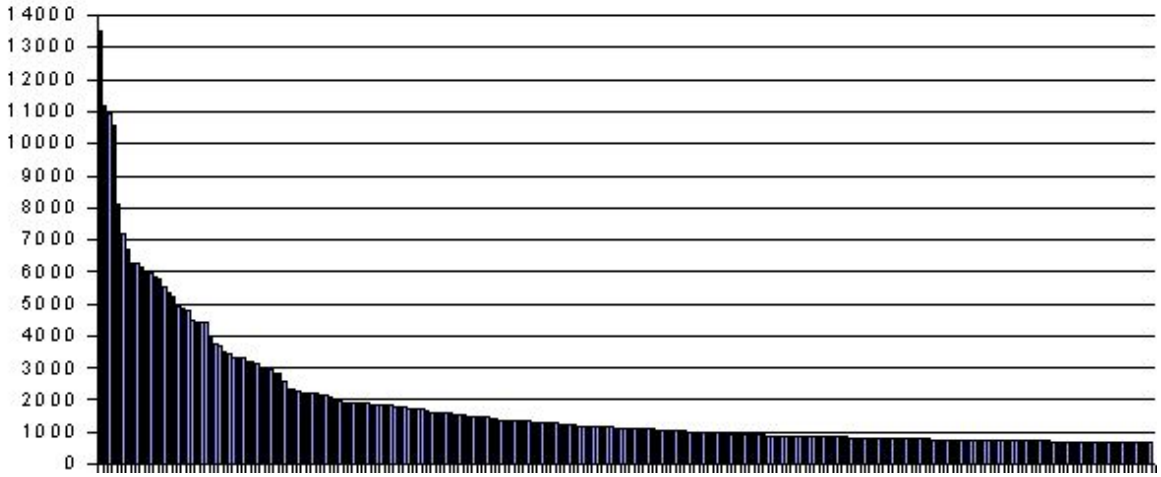
Merkmal/Korpus	K2003	K9704	RDeu	RSve	REngC	REng
Sprache	Deutsch	Deutsch	Deutsch	Schwedisch	Englisch	Englisch
Quelle	heise online	heise online	Reuters	Reuters	Reuters	Reuters
Domäne	IT+W	IT+W	EcoPol	Eco	Eco	EcoPol
Größe in MByte	17	94	170	20	525	1.323
Dokumente	8.861	48.501	116.212	15.732	379.158	806.791

2.2.5 Statistische Korpusdaten am Beispiel

Merkmal/Korpus	K2003	K9704	RDeu	RSve	REngC	REng
Worte	1.917.718	10.237.016	18.752.519	1.941.824	59.384.908	157.510.562
- verschiedene	143.746	435.042	353.949	73.111	412.058	744.077
Längstes Wort	40	48	40	33	26	26
Häufigste Wortlänge	3 (25,3%)	3 (25,3%)	3 (26,1%)	3 (22,9%)	3 (18,1%)	3 (18,6)
- verschiedene Worte	10 (8,6%)	11 (7,9%)	12 (8,0%)	8 (8,5%)	7 (15,5%)	8 (16,3)

Die Untersuchung des Zusammenhangs von Worthäufigkeit und Häufigkeitsrang bestätigte den von Zipf aufgedeckten Zusammenhang, wie Abbildung 11 zeigt.

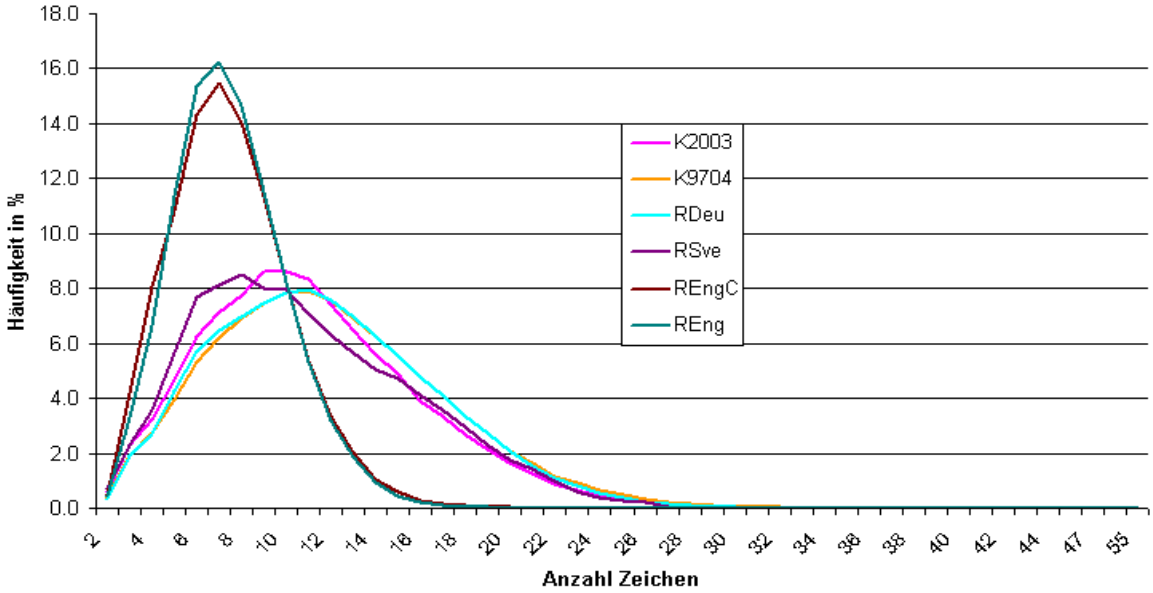
Abbildung 11: Zipf-Verteilung des Korpus K2003



Dargestellt sind aus Gründen der Skalierung nur die ersten 230 Worte der gesamten Rangfolge. Bemerkenswert ist, dass nur etwa 10% aller Worte mehr als 10 Mal auftreten. Betrachtet man den nach Luhn [Luhn58] relevanten Ausschnitt, so sollte dieser in etwa zwischen Rang 30 und 300, also bei h=6000 bis h=500 liegen. Damit fallen neben den zu seltenen auch häufige und unter Umständen relevante Worte wie ‚nicht', ‚Millionen' und ‚Euro' weg. Auf Seite 49 in Tabelle 4 wird dieser Aspekt vertieft. Trotzdem ist die damit erreichbare Volumenreduktion enorm und insbesondere verallgemeinernde Worte, die entsprechend häufiger genutzt werden, bleiben erhalten.

Im Zusammenhang mit der Häufigkeit einzelner Worte wurde für verschiedene Korpora untersucht, wie häufig Worte unterschiedlicher Länge auftreten. Die auf den ersten Blick überraschenden Ergebnisse stellt Abbildung 12 grafisch dar.

Abbildung 12: Wortlängenverteilung der betrachteten Sprachen



Die deutlich unterschiedliche Längenverteilung zwischen dem Deutschen und dem Schwedischen einerseits sowie dem Englischen andererseits erklärt sich jedoch aus der jeweiligen Art der Wortbildung. Im Englischen werden nur sehr selten Komposita gebildet, während dieses Vorgehen in den beiden anderen Sprachen sehr stark ausgeprägt ist. Hält man sich vor Augen, dass ein zusammengesetztes Wort mindestens zwei einzelne Worte zusammenfasst, erklärt sich auch das deutlich abweichende Verhältnis von insgesamt auftretenden zu paarweise verschiedenen Worten der einzelnen Sprachen.

Vor dem hier betrachteten Hintergrund der Texterschließung bedeutet dies, dass für Sprachen mit wenigen Komposita mehr Worte zur Bezeichnung eines Begriffs nötig sind. Daraus leitet sich ab, dass das Auftreten eines einzelnen Wortes in solchen Sprachen nur einen geringen Hinweis auf den damit auszudrückenden Begriff gibt. Dies führt wiederum zu einer höheren Komplexität bei der Segmentierung, also der Bestimmung, welche Worte zusammen gehören. Davon ausgehend ist zu vermuten, dass für Sprachen mit wenigen Komposita die syntaktische Analyse (POS-Tagging) wichtiger ist, weil dann auch grammatische Beziehungen zwischen den Worten als Klassifikationskriterium berücksichtigt werden können.

Da bei dem in dieser Arbeit verfolgten Ansatz kein Part-of-Speech-Tagging stattfinden soll, ist aus den in Abbildung 12 dargestellten Fakten zu erwarten, dass die Ergebnisse für die Englischen Korpora deutlich hinter denen für Deutsch oder Schwedisch zurück bleiben.

Dem Schwedischen und dem Englischen gemein ist, dass nur echte Eigennamen mit einem Großbuchstaben beginnen, wobei es im Englischen unter Umständen von der Intension des Verfassers abhängt, was als solcher betrachtet wird. Entsprechend ist die Schreibweise inkonsistent.

2.2.5.2 TF-IDF und Wahrscheinlichkeit

Mit Hilfe der Kenngröße TF-IDF sollen Indexterme, also Bezeichner zur Beschreibung des Inhalts von Dokumenten, ermittelt werden. Zur Erstellung eines Index mittels TF-IDF würden beispielsweise die 10 Worte mit den höchsten Werten pro Dokument angegeben.

In den Tabellen und sind die häufigsten Worte bzw. diejenigen mit dem höchsten TF-IDF-Wert angegeben, um den Zusammenhang zwischen diesen beiden Größen zu illustrieren. Bis Rang 19 sind alle Worte aufgeführt. Weiterhin wurden einige Worte angegeben, die subjektiv als wichtig für den Diskursbereich angesehen werden. Die Spalte ‚Max' gibt das häufigste Auftreten eines Wortes in einem Dokument wieder.

Die in aufgeführten Worte bis Rang 19 treten in sehr vielen Dokumenten auf und rufen daher nur kleine TF-IDF-Werte hervor. Häufig werden diese im Sinne einer Stoppwortliste zur Reduzierung des Textvolumens (siehe S.31) genutzt. Die 13 häufigsten Worte machen bereits 1/5 des gesamten Volumens des Korpus aus.

Bei einer inhaltlichen Erschließung der Texte ist die Verwendung einer Stoppwortliste ausgeschlossen. Wichtige Informationen gingen verloren, da beispielsweise Veränderungen oft durch Konstellationen wie ‚von ... auf' gekennzeichnet sind.

Die in den angeführten Worte größer Rang 19 stellen eine allgemeine Beschreibung des Diskursbereichs dar. Diese Worte kommen somit als Hyperonyme, also Oberbegriffe oder Generalisierungen in Frage, während höhere Werte für TF-IDF eher Eigennamen, also Begriffsinstanzen im eigentlichen Sinn (vgl. S.12),

darstellen. Kennzeichnend ist, dass diese tendenziell in weniger Dokumenten (Spalte ‚Docs') auftreten als allgemeinere Bezeichnungen.

Tabelle 2: Häufige und interessante Worte

Rang	Wort	Gesamt	Max	Docs	Max TFIDF
1	die	68455	66	8711	0,01707
2	der	64595	56	8753	0,01041
3	und	38705	39	8301	0,03858
4	in	27749	31	7889	0,05457
5	von	27312	29	7949	0,04772
6	das	24735	21	7681	0,04547
7	den	24051	20	7592	0,04684
8	mit	20655	24	7221	0,07442
9	für	19818	18	7082	0,06112
10	auf	18759	19	7067	0,06512
11	zu	17001	17	6624	0,07494
12	im	15136	15	6417	0,07334
13	des	14621	17	6420	0,08300
14	sich	13437	13	6184	0,07085
15	dem	13339	13	6181	0,07094
16	eine	12285	15	5914	0,09189
17	ein	12072	16	5814	0,10216
18	nicht	10776	15	5068	0,12698
19	auch	10036	13	5167	0,10624
44	Euro	5061	20	1993	0,45213
51	Unternehmen	4095	12	2604	0,22266
67	Microsoft	3112	15	1111	0,47191
77	Internet	2456	13	1488	0,35144
96	Kunden	1833	9	1232	0,26905
98	Software	1822	11	1197	0,33364
101	Deutschland	1810	10	1223	0,30005
116	Umsatz	1539	11	892	0,38266
125	Windows	1413	14	782	0,51494
136	USA	1294	7	929	0,23920
138	Markt	1273	6	1012	0,19725
141	IBM	1243	17	533	0,72402
149	Firma	1154	9	837	0,32176
160	SCO	1103	38	136	2,40480
170	Telekom	1003	14	408	0,65294

Tabelle 3: Bezeichnende und interessante Worte

Rang	Wort	Gesamt	Max	Docs	Max TFIDF
1	GHz	485	66	179	3,90203
2	SCO	1103	38	136	2,40480
3	Mio	420	30	69	2,20696
4	TRV	16	16	1	2,20349
5	TeraFlop/s	32	17	2	2,16268
6	Mosaic	18	17	2	2,16268
7	Kilby	15	15	1	2,06578
8	Athlon	260	25	95	1,71801
9	Bushnell	24	14	3	1,69502
10	Tablet	73	19	31	1,62808
11	Atanasoff	14	12	2	1,52659
12	Neumann	25	14	7	1,51529
13	ClamAV	11	11	1	1,51490
14	ITU	46	16	19	1,48969
15	nForce3	17	13	5	1,47333
16	Intertrust	14	12	3	1,45287
17	Palladium	85	16	26	1,41365
18	Batterien	26	14	12	1,40096
19	Sabre	13	11	2	1,39938
525	IBM	1243	17	533	0,72402
775	Telekom	1003	14	408	0,65294
1653	Windows	1413	14	782	0,51494
2084	Microsoft	3112	15	1111	0,47191
2283	Euro	5061	20	1993	0,45213
3591	Umsatz	1539	11	892	0,38266
4444	Internet	2456	13	1488	0,35144
4926	Software	1822	11	1197	0,33364
5227	Firma	1154	9	837	0,32176
5733	Deutschland	1810	10	1223	0,30005
9740	Kunden	1833	9	1232	0,26905
12785	USA	1294	7	929	0,23920
14297	Unternehmen	4095	12	2604	0,22266

legt weiterhin die Vermutung nahe, dass der Kenngröße TF-IDF auch ein temporaler Aspekt zukommt. Insbesondere bei der hier betrachteten, von häufigen Veränderungen gekennzeichneten Domäne ist zu erwarten, dass neue Entwicklungen in zeitlich beisammen liegenden Texten hohe TF-IDF-Werte erzielen. Gehen diese Entwicklungen in die allgemeine Begriffswelt der Domäne über, werden sie immer wieder erwähnt und erlangen Begriffsbedeutung. Verlieren die Entwicklungen ihre Bedeutung, so geht ihr TF-IDF-Wert mit der Zeit gegen 0. Diese Überlegungen werden bei der Interpretation von Tabelle 4 wieder aufgegriffen.

2.2.5 Statistische Korpusdaten am Beispiel

Zusammenfassend ist festzustellen, dass TF-IDF zumindest eine tendenzielle Aussage darüber gibt, ob ein Wort für ein Dokument eines Korpus relevant im Sinne von ‚Bedeutung tragend' ist. Worte mit mittleren Werten sind prinzipiell dazu geeignet, ein Dokument zu charakterisieren. Für eine sichere Beurteilung der darin behandelten Inhalte benötigt man jedoch eine Menge derartiger Worte. Nur so lässt sich die Auswirkung von Ausreißern egalisieren. Hat man eine Menge von Worten vorab als für ein Thema bezeichnend definiert, so lassen sich Dokumente anhand der Übereinstimmung beider Wortmengen klassifizieren. Für eine detailierte Erschließung ist TF-IDF jedoch nicht geeignet, da nur Worte, nicht aber deren Bedeutung betrachtet wird.

Im Rahmen des hier verfolgten Zieles, der begrifflichen Erschließung eignet sich die Kenngröße TF-IDF im Sinne einer Filterung. Die Überlegung dabei ist, dass insbesondere Worte mit einem hohen TF-IDF-Wert als Instanz eines Begriffs in Frage kommen und daher bei der sprachlichen Betrachtung zu berücksichtigen sind.

Auf Grund der bis hierher dargestellten Ergebnisse war zu vermuten, dass an Wahrscheinlichkeiten orientierte Maße auf Grund der betrachteten Mengen ungeeignet sind. Wie die Zipf-Verteilung zeigt, kommt die Mehrzahl der Worte weniger als 10 Mal vor und hat damit eine sehr geringe Auftretenswahrscheinlichkeit. Daher wurde der Korpus erst um das Jahresarchiv 2004 erweitert zu K0304 und anschließend auf die Jahre 1997 bis 2004 als K9704 ausgedehnt (Tabelle 4).

Um die Veränderung zu verdeutlichen, wurden aus den auftretenden Häufigkeitsklassen des K2003 je 3 Beispielworte gewählt und die entsprechenden Wahrscheinlichkeiten ermittelt. Diese blieben trotz zunehmendem Umfangs der Korpora nahezu unverändert.

Weiterhin sind die maximalen, auf das Intervall [0;1] normalisierten TF-IDF-Werte angegeben. Die farbige Hinterlegung dient der Verdeutlichung der Größenklassen. Schon dieser kleine, in Tabelle 4 dargestellte Ausschnitt der Korpora unterstützt die oben angestellte Vermutung hinsichtlich der temporalen Bedeutung von TF-IDF. Beispielsweise wurde ‚Flachdisplays' erst 2003 wichtig, während ‚MobilCom' in früheren Jahren (Öffnung des Telekommunikationsmarktes Ende der 90er Jahre) ein bezeichnender Term war.

Tabelle 4: Statistik mehrerer Korpora

Klasse (K2003)	Beispielwort	K2003	K0304	K9704	P(K2003)	P(K0304)	P(K9704)	TF(K2003)	TF(K0304)	TF(K9704)	Klasse (K9704)
Einmalig	Infrarotkamera		2	3		0,00000	0,00000	0,03252	0,03504	0,02490	Einmalig
84436 Worte	Spaghetti-Palast		1	1	0,00000	0,00000	0,00000	0,03252	0,03770	0,02773	257944 Worte
59%	Halbleiterspezialist		1	2		0,00000	0,00000	0,03252	0,03770	0,02595	59%
Zweimalig	Tasks		2	3		0,00000	0,00000	0,03256	0,03504	0,02490	Zweimalig
19576 Worte	Protokolldateien		4	7	0,00000	0,00000	0,00000	0,03256	0,03237	0,02273	57002 Worte
14%	Hemmschwelle		5	22		0,00000	0,00000	0,03256	0,03152	0,03981	13%
3 bis 5	Gegenspieler	4	7	16	0,00000	0,00000	0,00000	0,06197	0,06164	0,04153	3 bis 5
18037 Worte	Flachdisplays	5	7	21	0,00000	0,00000	0,00000	0,13022	0,12950	0,08307	52781 Worte
13%	Schlagwörter	3	4	10	0,00000	0,00000	0,00000	0,06511	0,06696	0,04477	12%
6 bis 10	WinHEC	8	14	37	0,00000	0,00000	0,00000	0,05540	0,05631	0,05749	6 bis 10
8507 Worte	Schublade	7	13	40	0,00000	0,00000	0,00000	0,02770	0,02785	0,01247	24886 Worte
6%	Host	6	17	64	0,00000	0,00000	0,00001	0,02830	0,05460	0,07049	6%
11 bis 100	Kreditkarten	22	38	93	0,00001	0,00001	0,00001	0,07277	0,07277	0,04940	11 bis 100
11227 Worte	Nutzerzahlen	11	26	49	0,00001	0,00001	0,00000	0,05263	0,05099	0,03566	34267 Worte
8%	Canon	70	147	245	0,00004	0,00004	0,00002	0,19850	0,19347	0,13824	8%
101 bis 1000	MobilCom	139	142	1100	0,00007	0,00004	0,00011	0,19849	0,19849	0,26644	101 bis 1000
1781 Worte	MByte	528	1139	2806	0,00028	0,00028	0,00027	0,19770	0,19035	0,12766	7098 Worte
1%	Intel	895	1893	6064	0,00047	0,00047	0,00059	0,21432	0,21495	0,13739	2%
1001 bis 10000	Microsoft	3098	6254	15901	0,00162	0,00156	0,00155	0,12150	0,17070	0,12662	1001 bis 10000
162 Worte	Auch	1864	3837	9817	0,00097	0,00096	0,00096	0,02663	0,03315	0,02249	960 Worte
0%	Telekom	1006	1888	7923	0,00052	0,00047	0,00077	0,16723	0,16878	0,12973	0%
10001 und öfter	nicht	10816	22523	59479	0,00564	0,00563	0,00581	0,03244	0,03216	0,07415	10001 und öfter
20 Worte	eine	11146	23428	60970	0,00581	0,00586	0,00596	0,02250	0,02254	0,04592	104 Worte
0%	der	64694	134316	345833	0,03373	0,03360	0,03378	0,00271	0,00311	0,00441	0%
143746	Gesamtwortzahl:	1917718	3997893	10237016							435042

Ein weiteres interessantes Ergebnis dieser Untersuchung ist, dass sich die Mächtigkeit der Häufigkeitsklassen zumindest bis zu dieser Menge kaum verändert. Auch in dem 93MByte umfassenden Gesamtkorpus mit über 435.000 verschiedenen Worten treten 90% seltener als 10 mal auf. Eine Erklärung dafür ist, dass insbesondere bei der zu Grunde liegenden Thematik die Namensentwicklung eine hohe Dynamik aufweist und stark von zusammengesetzten Worten geprägt ist.

Daraus ist zu folgern, dass zur Nutzung statistischer Methoden sehr umfangreiche Korpora nötig sind. Diese müssen bei volatilen Themengebieten aus einer kurzen Zeitspanne stammen um die Auswirkung zeitlicher Dynamik zu vermindern. Insbesondere bei sehr detaillierten Themengebieten, für die daher nur kleine thematisch zusammenhängende Korpora verfügbar sind, muss die aus der Grammatik der Sprache resultierende Vielzahl der Worte reduziert werden. Findet keine thematische Einschränkung statt stehen größere Korpora zur Verfügung. Allerdings wird dadurch die Auftretenswahrscheinlichkeit für Worte aus enthaltenen engen Themengebieten noch geringer.

Die angeführten Überlegungen dürften der Grund dafür sein, dass zum Vergleich von IR-Verfahren zur Textklassifikation oft Zeitungsarchive wie das von Reuters [Reut00] eingesetzt werden. Diese entstammen einem kurzen Zeitraum (hier ein Jahr) und weisen eine enorme Größe auf, sodass statistisch unterlegte Klassifikationsentscheidungen möglich sind.

2.2.5.3 Zusammengehörige Einheiten

Untersucht man den Korpus hinsichtlich Kollokationen, genauer 2-Grams, ergibt sich Tabelle 5, wobei offensichtlich ‚***Wall Street***' und ‚***Street Journal***' auch ein 3-Gram bilden könnten.

Bei den dargestellten Bezeichnern wird deutlich, dass die als Begriffsinstanz gekennzeichneten, sinnvollen 2-Grams zwischen etwa 30 und 300 mal im Korpus auftauchen. Die Auftretenshäufigkeit ist jedoch nur bedingt geeignet, da sie auch durch die Grammatik bedingte Kollokationen wie Artikel/Substantiv erfasst. Um diese Kombinationen zu eliminieren verwenden beispielsweise Biemann et al einen Wortartenfilter [Biem+03].

2.2.5 Statistische Korpusdaten am Beispiel

Tabelle 5: Kollokationsmaße im Vergleich

Begriff	2-Gram	Gesamt	1.Wort	2.Wort	DICE	Jaccard	GI
	Millionen Euro	1985	5643	5061	0,37089	0,22766	-0,03708
	Millionen US-Dollar	1600	5643	3931	0,33424	0,20065	-0,03452
	Milliarden US-Dollar	1094	2684	3931	0,33076	0,19815	-0,02543
	den USA	897	24051	1294	0,07078	0,03669	-0,08116
	Milliarden Euro	883	2684	5061	0,22802	0,12868	-0,03077
x	Mac OS	338	423	462	0,76384	0,61792	-0,00338
x	Red Hat	294	351	310	0,88956	0,80109	-0,00237
	von SCO	275	27312	1103	0,01936	0,00977	-0,09171
x	Wall Street	255	261	262	0,97514	0,95149	-0,00179
	Millionen Dollar	251	5643	676	0,07944	0,04136	-0,02710
	Nach Angaben	246	1824	1403	0,15246	0,08252	-0,01551
x	Street Journal	240	262	250	0,93750	0,88235	-0,00181
x	Internet Explorer	232	2456	245	0,17179	0,09397	-0,01238
x	Windows XP	215	1413	476	0,22763	0,12843	-0,00917
x	Deutsche Telekom	170	411	1003	0,24045	0,13666	-0,00712
x	Windows Server	163	1413	1142	0,12759	0,06814	-0,01290
x	Software AG	39	1822	547	0,03293	0,01674	-0,01258
	Microsoft Windows	35	3112	1413	0,01547	0,00780	-0,02242
x	AMD Athlon	34	450	260	0,09577	0,05030	-0,00429
x	Telekom Austria	33	1003	58	0,06221	0,03210	-0,00588
	Software Development	17	1822	80	0,01788	0,00902	-0,01000
	Microsoft Business	16	3112	192	0,00969	0,00487	-0,01623
	Microsofts Palladium	11	460	85	0,04037	0,02060	-0,00343
	Microsoft Internet	6	3112	2456	0,00216	0,00108	-0,02729
	Software Update	6	1822	294	0,00567	0,00284	-0,01142
	Voltage Athlon	5	16	260	0,03623	0,01845	-0,00183
	Unternehmen Software	3	4095	1822	0,00101	0,00051	-0,02845
	Unternehmen IBM	3	4095	1243	0,00112	0,00056	-0,02578
	Microsoft Mobility	3	3112	27	0,00191	0,00096	-0,01523
	Software Solutions	3	1822	169	0,00301	0,00151	-0,01070
	Software Symposium	3	1822	17	0,00326	0,00163	-0,00966
	AMDs Athlon	3	58	260	0,01887	0,00952	-0,00217
	Gates: Palladium	3	8	85	0,06452	0,03333	-0,00069

Sofern nicht beide Bestandteile sehr oft auch in anderen Verbindungen auftreten, scheint ein gutes Indiz für begriffliche Zusammengehörigkeit zu sein, wenn das gemeinsame Auftreten in etwa so häufig ist wie das des selteneren der beiden Worte. Dieser Aspekt liegt den beiden klassischen Maßen DICE und Jaccard zu Grunde, wobei hier qualitativ keine Unterschiede feststellbar sind[11].

Betrachtet man die gegenseitige Information (vgl. Definition VII, S.41) der Terme wie in Spalte ‚GI' dargestellt, erhält man deutlich bessere Ergebnisse in dem Sinne, dass nun bei absteigender Sortierung die Begriffsinstanzen dichter beisammen liegen. Die zwischen diesen Instanzen liegenden Bezeichner (z.B. ‚Software Symposium') kommen desweiteren zumindest prinzipiell ebenfalls als sinnvolle Einheiten in Betracht.

Als Ergebnis dieser sicher nicht erschöpfenden Untersuchungen zeigt sich, dass die Betrachtung der 2-Grams insbesondere unter Verwendung der Entropie, oft sinnvolle Ergebnisse liefert. Damit kommt diesem Maß eine ähnliche Bedeutung zu, wie die Größe TF-IDF für einzelne Terme. Zu beachten ist jedoch, dass die auf der Entropie basierende gegenseitige Information keinen Dokumentenbezug aufweist.

Kollokationen von mehr als zwei Worten sollen nicht beachtet werden. Eine Möglichkeit diese zu identifizieren wäre innerhalb stark verbundener 2-Grams nach Verbindungsmöglichkeiten wie beispielsweise ‚Wall Street' und ‚Street Journal' zu suchen.

Abschließend zu den Methoden des klassischen Information Retrieval wird im Folgenden das Vektorraummodell vorgestellt und damit die Problematik geeigneter Repräsentationsformen eingeleitet.

11 Eine absteigende Sortierung ruft bei beiden Größen die gleiche Reihenfolge hervor.

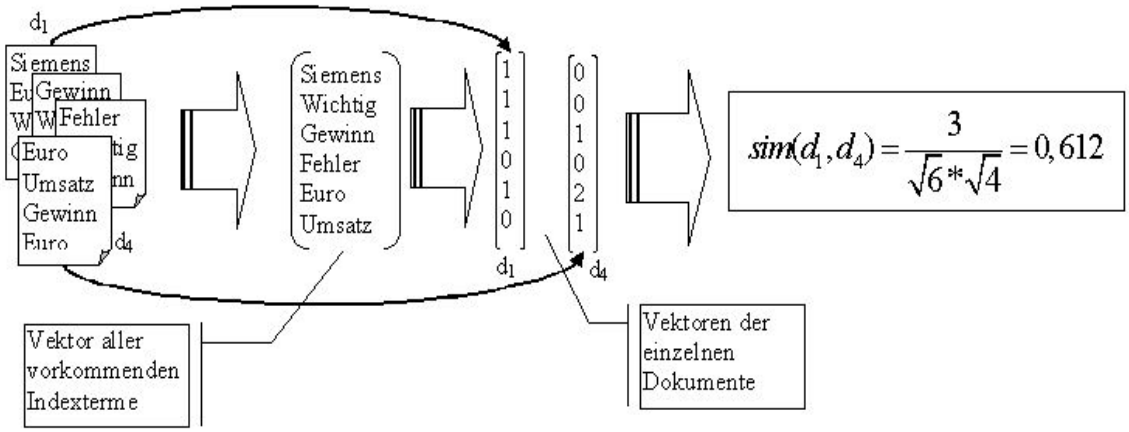
2.3 Repräsentationsformen

Nach der Vorstellung des klassischen Vektorraummodells widmen sich die folgenden Kapitel der symbolischen Repräsentation mit Hilfe propositionaler und regelbasierter Methoden.

2.3.1 Vektorraummodell

Dieses Modell zur Repräsentation von Dokumenten wurde 1983 von Salton und McGill [SaMcG83] vorgestellt und ist die bis heute am häufigsten genutzte Repräsentationsform. In Abbildung 13 ist die prinzipielle Vorgehensweise am Beispiel illustriert. An dieser Stelle soll nur das Vorgehen, nicht jedoch mögliche Verbesserungen vorgestellt werden.

Abbildung 13: Funktionsprinzip des Vektorraummodells



Quelle: Eigene Erstellung

Ausgangspunkt ist eine Textsammlung, der eine Menge von Indextermen zugeordnet ist. Diese Indexterme werden in Form eines Vektors organisiert, so dass jede Dimension dieses Vektors einem Indexterm entspricht. Anders ausgedrückt: Durch diesen Vektor ist ein kontrolliertes Vokabular zur Beschreibung von Dokumenten definiert und den Dimensionen eine Semantik zugeordnet, so dass jede Dimension einen bestimmten Term repräsentiert. Damit bildet dieser Vektor den Prototyp weiterer Vektoren, die der Beschreibung der einzelnen Dokumente dienen.

Für jedes Dokument wird ebenfalls ein solcher Vektor angelegt. Für jede Dimension, also jeden möglichen Indexterm wird dann untersucht, ob und mit welchem Gewicht dieser Term in diesem Dokument auftritt. Dieses wird an die entsprechende

Position des Dokumentenvektors eingetragen. Im einfachsten Fall, wie in Abbildung 13 dargestellt, ist dies die Häufigkeit dieses Terms oder dessen TF-IDF-Wert.

Die Ähnlichkeit zwischen zwei Dokumenten lässt sich so auf den Abstand zweier n-dimensionaler Vektoren zurückführen. Als Ähnlichkeitsmaß wird üblicherweise der Kosinus des Winkels zwischen diesen Vektoren gemäß Formel 9 bestimmt. Abbildung 13 verdeutlicht auch dies.

Formel 9: Kosinus-Ähnlichkeit zweier Vektoren

$$sim(x,y) = \cos\alpha = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} * \sqrt{\sum_{i=1}^{n} y_i^2}}$$

Damit stellt dieses Modell eine einfache und effiziente Repräsentationsform dar, die stark von der zunehmenden Leistung der Rechentechnik profitiert. Ein weiterer Vorteil ist die leichte Erweiterbarkeit des Vektorraummodells. So haben Hotho et al [HoSt+03] eine kategorisierende, begriffliche Komponente aufgenommen und konnten so die Retrievalqualität verbessern.

Im Folgenden werden weitere Repräsentationsmöglichkeiten vorgestellt, soweit sie für den hier verfolgten Ansatz interessant sind. Dies sind in erster Linie deklarative Formen, die der Beschreibung von Sachverhalten dienen [KarTe01, S.54]. Im Gegensatz zur Repräsentation von Dokumenten im Vektorraummodell lassen sich diese Sachverhalte bezüglich ihres Wahrheitswertes beurteilen.

2.3.2 Symbole und Logik

Grundlage deklarativer Repräsentationen ist die Aussagenlogik. Da diese Thematik an dieser Stelle nur umrissen werden kann, sei ansonsten auf die entsprechende Literatur wie [Stein+64] oder [RuNo04, S.260ff] verwiesen.

Eine Aussage sei in Anlehnung an Steiner [Stein+64, S.197] definiert als:

Definition VIII: Aussage

> *Eine Aussage ist ein sprachliches Gebilde, für das es sinnvoll ist zu fragen, ob es entweder wahr oder falsch ist.*

2.3.2 Symbole und Logik

Die Darstellung der Aussagen erfolgt mit Hilfe von Symbolen. Eine einzelne, nicht weiter zerlegbare Aussage wird als atomar bezeichnet. Komplexe Aussagen entstehen durch Verknüpfung von Symbolen zu Symbolstrukturen mit Hilfe der Junktoren ‚UND', ‚ODER', ‚NICHT' sowie darauf aufbauend ‚IMPLIZIERT'.

Die sprachlichen Äußerungen der Aussagenlogik sind folglich auf Verknüpfungen elementarer Aussagen durch Junktoren beschränkt und damit ungeeignet für analytische Zwecke. So kann ein Satz wie ‚Die Katze sitzt auf dem Dach.' aussagenlogisch nur als atomares Gebilde behandelt werden, welches als wahr oder falsch klassifiziert werden kann. Die durch den Satz ausgedrückte Relation zwischen Katze und Dach kann erst durch die mächtigere Prädikatenlogik dargestellt werden.

Ein Prädikat bezeichnet eine ein- oder mehrstellige Relation auf einer Menge [Stein+64, S.210]. Einstellige Prädikate weisen damit einem Element eine Eigenschaft zu und mehrstellige Prädikate setzen zwei oder mehr Elemente zueinander in Bezug. Im Beispiel würde man notieren sitzt_auf(Katze, Dach). Das Prädikat ‚sitzt_auf' verbindet demnach die Elemente ‚Katze' und ‚Dach'. Diese Form wird als Prädikat-Argument-Struktur bezeichnet. Insbesondere einstellige Prädikate können als Zusprechung von Eigenschaften zu einem Objekt, also Prädikation im umgangssprachlichen Sinn, verstanden werden [Wede92]. Entsprechend würde sitzt_auf_Dach(Katze) notiert um deutlich zu machen, dass die Katze die Eigenschaft hat, auf dem Dach zu sitzen.

Der Übergang zur Prädikatenlogik bringt neben der Möglichkeit zur Abbildung von Beziehungen auch den Vorteil, dass Objekte, also abstrakte Einheiten, verknüpft werden können. An die Stelle der Symbole treten damit Variable und man spricht von ***Aussageformen*** statt von Aussagen. Für Aussageformen wiederum werden Quantoren benötigt um auszudrücken, ob ein Prädikat für genau ein, mindestens ein oder für alle Elemente einer Menge wahr ist. In der Syntax der Prädikatenlogiklogik sieht das Beispiel dann wie folgt aus:

$$\exists x \exists y (Katze(x) \wedge Dach(y)) \wedge sitzt\_auf(x,y)$$

bzw. eher umgangssprachlich notiert:

$$\exists x \in \{Katzen\} \wedge \exists y \in \{Dächer\} : sitzt\_auf(x,y)$$

Die Anwendung der Logik zur Darstellung sprachlicher Konstrukte geht zurück auf die Kognitionsforschung nach Pylyshyn [Pyly73]. Die Aussagen werden hier als Propositionen bezeichnet und geben den deklarativen, also auf den Wahrheitswert bezogenen Gehalt einer sprachlichen Äußerung wieder [Kluw92, S.153]. Die Repräsentation sprachlicher Äußerungen mit Hilfe der Logik wird daher auch als ***propositional*** oder ***deklarativ*** bezeichnet, wobei eine Proposition nach Sowa [Sowa84, S.139] die Darstellung einer Aussage ist[12].

Die Repräsentation in dieser Form hat zwei große Vorteile: einerseits ist sie leicht verständlich und für natürlichsprachige Konstrukte anwendbar und andererseits lässt sie sich mit Hilfe der Prädikatenlogik auch maschinell verarbeiten. Die Möglichkeit Variablen zu verwenden entspricht unmittelbar dem Verständnis vom Begriff, da damit vom konkreten Element, dem Sachverhalt, auf eine Klasse von Elementen, den Begriff, abstrahiert wird. Durch die Nutzung von Prädikaten werden Objekten Eigenschaften zugewiesen und in Zusammenhang gebracht, was der Abbildung des Begriffsverständnisses (vgl. Abschnitt 2.1.1) entspricht. Tatsächlich geht die Prädikatenlogik zu großen Teilen auf die sprachphilosophischen Überlegungen Freges in seiner Begriffsschrift [Freg86] zurück.

Die Einschränkung auf zwei Wahrheitswerte ist für den angestrebten Einsatz ohne Bedeutung, da Aussagen wie ‚Manchmal regnet es.' nicht Gegenstand der Betrachtungen sind. Die hier zu betrachtende Logik repräsentiert Aussagen unabhängig von der Zeit. Die Einbeziehung temporaler Aspekte ist als Erweiterung zu sehen und daher je nach Anwendung beispielsweise durch Angabe von Zeitpunkten oder Zeiträumen zu implementieren. Diese Erweiterungen sind Gegenstand der Modallogik [HuCr78].

Ebenfalls auf Implementationsebene zu behandeln ist die Frage der Effizienz. Zur Verifikation einer Aussage muss in den beteiligten Mengen gesucht werden, ob eine Konstellation auftritt, die alle Bedingungen erfüllt. Hier sind Mechanismen nötig, die die Suche einschränken können. Für die inhaltliche Erschließung von Texten erscheint diese Problematik aber ebenfalls zweitrangig, da keine zu beweisenden Behauptungen postuliert sondern lediglich Fakten abgebildet werden.

12 Entsprechend der Bedeutung des Wortes; vgl. Merriam Webster: http://www.webster.com/cgi-bin/dictionary?sourceid=Mozilla-search&va=proposition

Die Möglichkeiten logischer Repräsentation sind damit als geeignet für eine begriffsorientierte Vorgehensweise im Information Retrieval zu bewerten und es stellt sich die Frage, wie diese Repräsentation organisiert, also tatsächlich umgesetzt wird. Dazu werden im Folgenden darauf aufbauende Formalismen vorgestellt.

2.3.2.1 Frames

Aufbauend auf die Arbeiten von Pylyshyn [Pyly73] entwickelte Minsky zur Beschreibung von Wahrnehmungsszenen das Konzept der Frames [Minsk75]. Dabei handelt es sich um eine komplexe Datenstruktur, die ein Objekt eines Problembereichs prototypisch beschreibt.

Ein Frame entspricht einer Klasse eines Problembereichs, die zu anderen Frames in Beziehung stehen kann, wobei in hierarchischen Beziehungen Eigenschaften vererbt werden können. Diese Eigenschaften werden als Slots bezeichnet und können mit verschiedenen Ausprägungen belegt werden [KarTe01, S.64]. Den Slots können Prozeduren zugeordnet sein die beschreiben, welche Aktionen bei Änderung, Einfügung, Abfrage oder Löschung einer Ausprägung auszuführen sind. Weiterhin können für diese Slots ***Constraints***, also Einschränkungen bezüglich des Wertebereichs definiert werden. Sind alle Slots eines Frame mit den konkreten Ausprägungen eines Objekts belegt, so spricht man von einer Instanz.

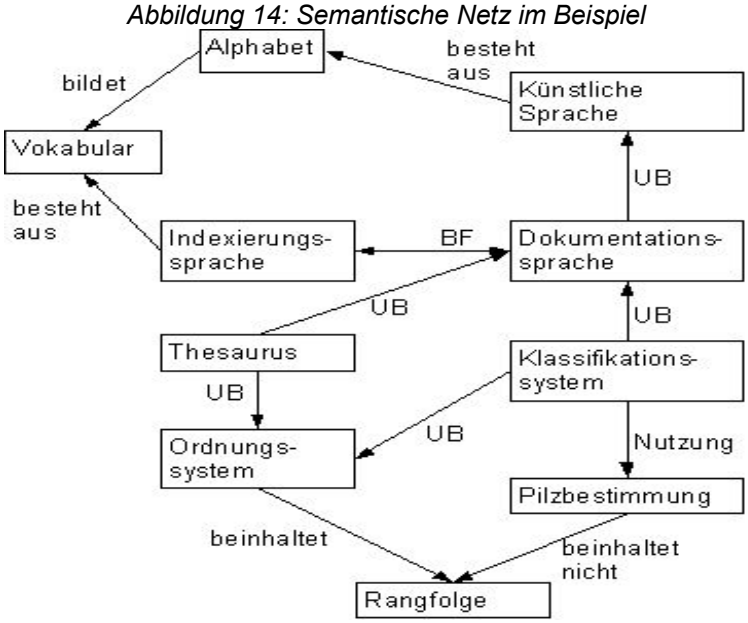
Problematisch sind Frames, falls reale Objekte stark vom definierten Prototyp, also der Klasse, abweichen oder überhaupt nicht definiert sind. Daher eignen sich Frames in erster Linie für statische und thematisch eingeschränkte Problembereiche, so dass vorab definiert werden kann, welche Objekte von Interesse sind.

2.3.2.2 Semantische Netze

Ebenso wie Frames stammen semantische Netze aus der Psychologie und dienten der Modellierung des menschlichen Gedächtnisses [KarTe01, S.77]. Das Augenmerk liegt dabei auf den Beziehungen zwischen Objekten, die ihrerseits beispielsweise durch Frames dargestellt werden können. Diese Beziehungen werden als gerichtete und beschriftete Kanten dargestellt.

Ein einfaches Beispiel zeigt Abbildung 14 in Anlehnung an den Thesaurus von Abbildung 9. Während dort lediglich Vererbungen auftraten, bieten die neuen Verbindungen mehr Beschreibungsmöglichkeiten. Ein semantisches Netz erlaubt damit eine einfache, leicht nachvollziehbare Notation eines Sachverhalts. Da jedoch keine standardisierte Terminologie existiert, muss bei der Anwendung Konsens darüber bestehen, was unter einer Bezeichnung zu verstehen ist.

Abbildung 14: Semantische Netz im Beispiel



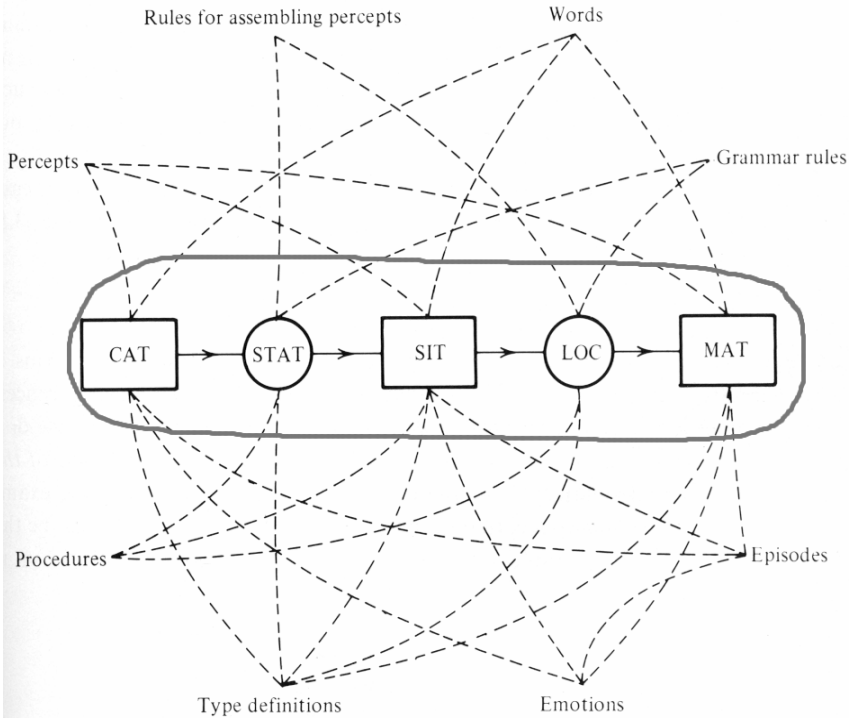
Quelle: eigene Erstellung

Dies macht auch die Schwierigkeit semantischer Netze deutlich: sie bilden stets eine spezielle Sichtweise ab. So könnte die Verbindung zwischen Alphabet und Vokabular auch entgegengesetzt gerichtet sein (‚wird gebildet aus').

Weiterhin wird nicht zwischen Begriffen und Sachverhalten unterschieden. So könnte ‚Alphabet' in Abbildung 14 durchaus ein konkretes Alphabet bezeichnen. Dies verhindert eine eindeutige Interpretation eines semantischen Netzes.

Sowa [Sowa84, S.76] definiert ein semantisches Netz als die Menge aller Verbindungen zwischen Begriffen und anderen Begriffen, Wahrnehmungen, Vorgängen, etc. Die Verbindung zwischen den Elementen einer Aussage (Proposition) im Sinne der Prädikatenlogik wird hingegen als ***Conceptual Graph***, in Abbildung 15 hervorgehoben, bezeichnet. Dieser entspricht jedoch der allgemeinen Vorstellung eines semantischen Netzes [Sowa84, S.78; KarTe01, S.78ff].

Abbildung 15: Conceptual Graph und semantisches Netz



Quelle: [Sowa84, S.77]

2.3.2.3 Ontologien

Um über einen Sachverhalt zu kommunizieren ist es essentiell, eine eindeutige Beschreibung des betreffenden Diskursbereichs (Domäne) bei allen Beteiligten vorzufinden [AuWe02]. Die innerhalb dieses Bereichs auftretenden Worte müssen inhaltlich eindeutig bestimmt und in klar definierter Relation zueinander stehen, damit sichergestellt ist, dass jede Bezeichnung stets das Gleiche meint [Mönc03].

Eine solche explizite Beschreibung einer Begriffswelt wird als Ontologie bezeichnet [Sowa84, S.294; Grub93, S.1]. Dementsprechend sind alle für einen bestimmten Diskursbereich definierten Vokabulare Ontologien, unabhängig davon, ob es sich dabei um Listen von Schlüsselworten[13], Thesauren oder semantische Netze handelt [Lehn06, S.195f]. Die Bezeichnung ‚Ontologie', also Lehre vom Seienden, deutet diese sehr allgemeine Auffassung bereits an, die zu einer Vielzahl von Interpretationen führt (z.B. [QuWo02], [Maed+03], [Hand+03]). Diese Bezeichnung erlaubt so einen nicht weiter spezifizierten Verweis auf eine Begriffssystematik.

13 Im Sinne thematischer Beschreibungen; vgl. S. 34.

Letztendlich geht es um ein Modell einer Domäne, das die interessierenden Objekte sowie deren Beziehungen zueinander in formalisierter, also nicht natürlichsprachiger Form wiedergibt [LoKa99].

Darüberhinaus weisen Ontologien Möglichkeiten zur Repräsentation von Existenz von Vererbungs- und Existenz- sowie Gültigkeitsregeln ***(Constraints)*** auf. Damit wird eine logikbasierte Repräsentation ermöglicht, so dass auf begrifflicher Ebene definierte Eigenschaften als Grundlage für logische Schlüsse auf Ebene von Begriffsinstanzen möglich sind. Auf Grundlage dieser Beschreibungslogik wurde die ***Web Ontology Language OWL***, ein Standard zur Beschreibung von Ontologien im ***Semantic Web*** entwickelt [W3C01, W3C04]. Die Mehrzahl der verfügbaren Ontologien beschränkt sich jedoch auf einfache Hierarchien von Begriffen und geringe Detaillierungsgrade [Noy05].

2.3.2.4 Generierung von Ontologien

Um die Modellierung eines Diskursbereichs zu unterstützen finden häufig Kollokationsanalysen (vgl. Seite 35) Verwendung. Dabei wird an einem thematisch eingegrenzten Textkorpus untersucht, wie häufig Begriffe innerhalb eines gewissen Abstands zueinander vorkommen [QuWo02]. In einem Text sind jedoch lediglich Bezeichner von Begriffen oder Begriffsinstanzen zu finden. Daher werden vorab, beispielsweise mit Hilfe eines Wörterbuchs, Synonyme auf eine gemeinsame Hauptbezeichnung abgebildet, die dann mit dem Begriff gleichgesetzt wird [HaSc98].

Zur Identifikation hierarchischer Beziehungen wurden in frühen Arbeiten manuell syntaktische Konstrukte identifiziert, wie sie in der natürlichen Sprache benutzt werden [Hear92]. Das Prinzip sei an folgendem Beispiel verdeutlicht: „Fluor, Chlor, Brom und Jod sind Halogenide.“. Überführt man diesen Satz in sein syntaktisches Muster[14] entsteht „NP{, NP}+ und NP sind NP.“, womit die NP links von ‚sind‘ als Unterordnung der NP auf der rechten Seite identifiziert werden können. Die Hauptbezeichnung ist in diesem Fall ein Wort der Art NP rechts von ‚sind‘.

14 In der Notation regulärer Ausdrücke; NP - Nominalphrase

2.3.2 Symbole und Logik

Neben dem enormen Aufwand zur Identifikation und Erstellung solcher Muster bereitet die fehlende sprachliche Flexibilität die größten Schwierigkeiten, denn selten sind natürlichsprachige Sätze derart simpel aufgebaut. Ein weiteres Problem in diesem Zusammenhang ist die schwer zu lösende Ambiguität. Gegeben seien die Sätze „Klaus und Dieter sind Kinder.“, „Klaus und Dieter sind Vornamen.“ und „Klaus und Dieter sind Schüler.“ Ohne weitere Information würde man schließen, ‚Kinder‘, ‚Vornamen‘ und ‚Schüler‘ seien synonym.

Einen Ansatz zur Unterstützung der Erstellung syntaktischer Muster mit Hilfe von Clusteranalysen stellten Faure und Nedellec [FaNe98] mit ASIUM vor, welches Vorschläge für derartige Muster generiert. Um dem Problem der Ambiguität zu begegnen wird hier neben der Identifikation der Hyponyme[15] untersucht, ob diese in mindestens zwei Fällen in Verbindung mit dem gleichen Verb auftreten. Würde im Beispiel weiter im Text „...Kinder gehen...“ und „...gehen Schüler...“ stehen, so wäre dies als Indiz dafür zu sehen, dass ‚Kinder‘ und ‚Schüler‘ tatsächlich synonym sind.

Cimiano et al [CiHo+05] nutzen aufbauend auf eine syntaktische Analyse die Verbindung zwischen Subjekt und Objekt durch ein Verb sowie die modale bzw. kausale Beschreibung von Sachverhalten durch Präpositionen (‚infolge‘, ‚einschließlich‘, etc.). Die Zuordnung zwischen diesen Satzteilen wird aus dem bei der syntaktischen Analyse erstellten Syntaxbaum (vgl. Abbildung 6, S.18) gewonnen. Mit Hilfe probabilistischer Verfahren (vgl. Abschnitt 2.2.4) wird dann das gemeinsame Auftreten der Worte untersucht und diese entsprechend hierarchisch aggregiert.

Gemeinsame Voraussetzung aller dieser Verfahren ist die Verwendung eines POS-Taggers sowie eine thematische Eingrenzung der zu Grunde liegenden Texte, so dass die Bedeutung der Worte eindeutig ist (vgl. [HaSc98, FaNe98]).

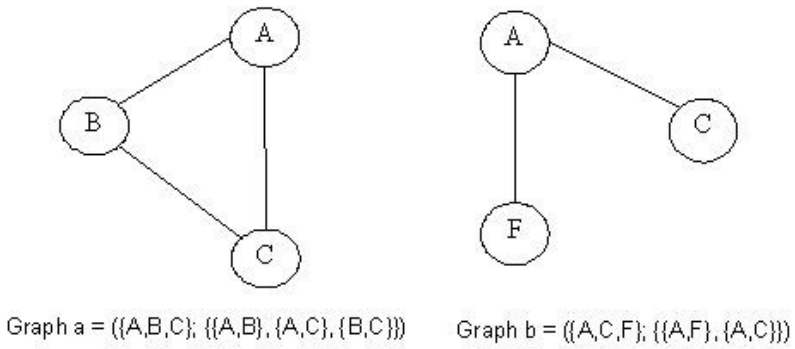
Mit Hilfe von Ontologien repräsentierte Diskursbereiche nutzen ebenso wie die strukturale Textanalyse zur inhaltlichen Interpretation von Texten [Titz93] strukturelle Darstellungsformen. Diese haben ihr Fundament in der Graphentheorie, wobei im Folgenden nur Kerngedanken wiedergegeben werden sollen. Für tiefer gehende Aspekte sei auf weiterführende Literatur wie [Titt03] verwiesen.

15 Unterbegriff; Gegenstück: Hyperonym / Oberbegriff.

2.3.2.5 Struktur und Graphen

Ein ungerichteter Graph G=(V,E) besteht aus nicht leeren Mengen von Knoten V und Kanten E, wobei jeder Kante e ∈ E zwei Knoten v ∈ V zugeordnet sind. Entsprechend lässt sich eine Kante e durch e = {u,v} mit u,v ∈ V beschreiben, wobei u und v als Endknoten der Kante e bezeichnet werden (siehe Abbildung 16). Der ***Grad*** eines Knotens entspricht der Anzahl Kanten, an denen er beteiligt ist. Knoten, die nicht durch Kanten mit anderen Knoten verbunden sind, werden als isolierte Knoten bezeichnet.

Abbildung 16: Zeichnerische und Mengennotation von Graphen



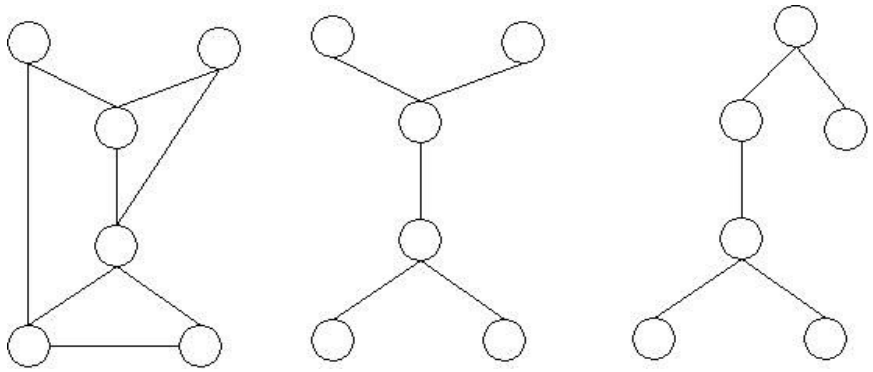
Quelle: eigene Erstellung

Bei gerichteten Graphen wird den Kanten jeweils ein Richtungssinn zugeordnet, so dass der Reihenfolge der Notation der Endpunkte eine Bedeutung zukommt. Dies wird durch die Notation in Form von Tupeln verdeutlicht. Die Kante (u,v) ist der Kante (v,u) entgegen gerichtet, so dass der Graph nur in Richtung der Kante durchlaufen werden kann. Der Knoten am Ursprung der Kante wird dann als Startknoten bezeichnet. Gewichtungen der Kanten dienen dazu, semantische Ähnlichkeiten zu definieren.

Ein zusammenhängender, kreisfreier Graph[16] wird als Baum bezeichnet und hat m Knoten sowie m-1 Kanten. Knoten vom Grad 1, die nur Endknoten sind, werden als Blätter bezeichnet. Existiert ein Knoten, der nur Startknoten ist, so wird dieser als Wurzel und der Baum entsprechend als Wurzelbaum bezeichnet. Beispiele verschiedener Graphen zeigt Abbildung 17.

16 Alle Knoten des Graphen sind über Kanten erreichbar und es existiert kein Weg durch den Graphen, bei dem eine Kante zweimal durchlaufen wird.

Abbildung 17: Graph, Baum und Wurzelbaum



Quelle: eigene Erstellung

Auf Grund dieser Eigenschaften eignen sich Wurzelbäume insbesondere zur Darstellung hierarchischer Zusammenhänge und Taxonomien, wie sie für Begriffssysteme verwendet werden. Um dafür geeignete Ähnlichkeitsmaße definieren zu können, wird später auf Maße der strukturellen Ähnlichkeit von Graphen eingegangen.

2.3.3 Bezug zur Arbeit

Ziel des Abschnitts 2.3 ist es, geeignete Formen der Repräsentation für ein am Begriff orientiertes Information Retrieval aufzuzeigen. Das dieser Arbeit zu Grunde liegende Begriffssystem soll dazu dienen, die zu betrachtenden Begriffe sowie deren Zusammenhang zwischen einander zu definieren. Ein weiterer Aspekt ist es, die mögliche Einbindung in ein größeres Umfeld zu illustrieren. Die dabei zum tragen kommenden Fragen der strukturellen Ähnlichkeit bei der grafischen Repräsentation von Begriffssystemen sollen jedoch nur schematisch verfolgt werden.

Der Domänenbezug einer Ontologie bringt mit sich, dass darauf aufbauende Informationssysteme nur für ein konkretes Wissensgebiet ausgelegt sind. Für praktische Anwendungen stellt dies aber keinen Nachteil dar. Hier sind die zu Grunde liegenden Informationsmengen ohnehin thematisch eingeschränkt.

Vor diesem Hintergrund erscheint es sinnvoll, zur Repräsentation ein semantisches Netz als ontologisches Modell zu wählen. Innerhalb der darzustellenden Objekte interessieren unter Umständen verschiedene Attribute wie Zeitpunkte, Kommentare oder Quantitäten. Daher sind an dieser Stelle Frames als geeignet anzusehen.

Auf Grund des prototypischen Charakters der Arbeit ist es angemessen, die Ontologie manuell zu erstellen und nicht zu generieren, zumal dies vom eigentlichen Fokus ablenken würde. Trotzdem ist zu erwarten, dass Ansätze deutlich werden, die zu einer Erweiterung bzw. Detaillierung der Ontologie führen. Diese sind zu diskutieren und bei vertretbarem Aufwand weiter zu untersuchen. Dabei liefert das explizit gemachte Modell auf Grund der Verbindungen zwischen den Begriffen unter Umständen wichtige Hinweise zur Plausibilität. Stehen Begriffe gemäß Begriffssystem in einer engen Beziehung, so sollte sich dies in einem Dokument dieser Domäne widerspiegeln.

Die Herausforderung besteht jedoch darin, die Ontologie mit ‚Leben', also den Instanzen der beschriebenen Begriffe, zu füllen [Var+02]. Daher liegt der Schwerpunkt in der Identifikation und Extraktion der Informationen und wird im folgenden Kapitel genauer dargestellt. Erst damit lassen sich die Aussagen eines Textdokuments darstellen und verarbeiten. Die Einordnung der extrahierten Fakten in ein Begriffssystem entspricht einerseits deren Klassifikation und andererseits deren Vernetzung durch das Modell. Da eine Begriffswelt jedoch nicht nur wahrnehmbare Objekte enthält ist es notwendig, auch abstrakte Begriffe identifizieren zu können.

2.4 Inhaltlich- begriffliche Erschließung

Logikbasierte Repräsentationen also sind geeignet, vorhandene Informationen in ihrem semantischen Zusammenhang abzubilden und ermöglichen einen schnellen und gezielten Zugriff darauf. Aus Sicht des Information Retrieval stellt sich nun die Frage, wie diese Informationen mit möglichst geringem Aufwand aus Texten gewonnen werden können. Das Ziel ist die semantisch konsistente Beschreibung der in einem Textdokument behandelten Inhalte [StHa03]. Daher gilt es, die Inhalte zu identifizieren und als Instanzen der in der Ontologie definierten Objekte zu klassifizieren.

Die folgenden Abschnitte widmen sich jedoch nicht nur der eigentlichen Informationsextraktion sondern gehen zuvor der Frage nach, wie sich Inhalte von Dokumenten auf übergeordneter, thematischer Ebene erschließen lassen. Dies ist einerseits zur Zuordnung der Dokumente zum betroffenen Diskursbereich und somit der

passenden Ontologie notwendig. Andererseits sollten die hierzu verwendeten Verfahren auch zur Anwendung auf Ebene der einzelnen Zeichenfolgen genutzt werden können. Nach einem Exkurs zu konnektionistischen Methoden endet dieses Kapitel mit einem Blick auf ergänzende Verfahren, die an den Bereich der inhaltlichen Erschließung angrenzen.

2.4.1 Metadaten

Unter Metadaten werden Daten über Daten, in diesem Fall über Dokumente, verstanden. Dies sind beispielsweise Angaben zum Autor, dem Erscheinungszeitpunkt oder dem behandelten Diskursbereich. Diese Angaben ermöglichen dem Nutzer eines Informationssystems eine erste Bewertung der möglichen Relevanz eines Dokuments, da sie als Strukturierungsmerkmale die Einordnung in bekannte Zusammenhänge ermöglichen [DCore06].

Zur Realisierung dieser Auszeichnungen haben sich Auszeichnungssprachen etabliert, deren bekanntester Vertreter die ***eXtensible Mark-up Language*** (XML) ist. Im Bereich des Internet wurde eigens ein META-Tag eingeführt, um derartige Informationen aufnehmen zu können. Die Angaben zum Inhalt der Dokumente sind jedoch manuell erstellt und daher mit Vorsicht zu interpretieren. Sie werden oft genutzt, um Suchmaschinen gezielt zu manipulieren [Bage05]. Aus diesem Grund wird versucht, Inhaltsangaben über Häufigkeiten von Schlüsselworten oder aus Strukturierungsmerkmalen zu generieren.

Verfahren in diesem Umfeld nutzen Merkmale wie Überschriften und Besonderheiten in der Formatierung wie Großschreibung, Font, Schriftgröße, Stil und Absätze [GoAm00, S.49]. Bei Dokumenten, die in einer Auszeichnungssprache kodiert sind, lassen sich diese Elemente an Hand der entsprechenden HTML- bzw. XML-Tags identifizieren und damit maschinell erfassen.

2.4.2 Klassifikation und Clustering

Die Kenntnis des Diskursbereichs eines Korpus bzw. der darin enthaltenen Dokumente ist für eine automatische Analyse oft von Vorteil, beispielsweise zur Auswahl des geeigneten Vokabulars zur Indexierung. Außerdem gestattet diese Kenntnis ein

effizienteres Retrieval, da nicht interessierende Themen von der Suche ausgeschlossen werden können [CiHo+05].

Ein Diskursbereich gliedert sich im Allgemeinen in mehrere Unterbereiche, wie beispielsweise zu ‚Informatik' ‚Hardware' und ‚Software' gehören. Je nach thematischer Eingrenzung des Korpus lassen sich so mehrere Detailierungsstufen bzw. Einzelthemen identifizieren [Gant05]. Führt man diese Untergliederung weiter, so gelangt man auf die Ebene einzelner Begriffe und deren Instanzen [Stum+01].

Weist man einem Dokument einen Bezeichner eines Diskursbereichs zu, so entspricht dies einer Kategorisierung [Strö04]. Dadurch lassen sich auf der jeweiligen Ebene thematisch ähnliche Dokumente oder Dokumentteile zusammenfassen. Sind die in Frage kommenden Kategorien vorab festgelegt, spricht man von Klassifikation, anderenfalls von Clustering. Von Interesse sind an dieser Stelle lediglich maschinelle Verfahren, wobei zuerst auf die Klassifikation eingegangen werden soll.

Die Anpassung des Klassifikators zur möglichst exakten Klassifikation anhand einer Trainingsmenge wird als überwachtes Lernen oder ***supervised learning*** bezeichnet [MaHü+03, S.97ff]. Das Lernen besteht dabei darin, für jede Klasse Merkmale sowie maximal zulässige Abweichungen davon zu identifizieren, die ein Element haben muss, um dieser Klasse zugeordnet zu werden. Bei der Zuordnung eines Elements zu einer Klasse handelt es sich damit um eine binäre Entscheidung.

Da beim Clustering nicht bekannt ist, welche Kategorien existieren, kann nicht anhand einer Trainingsmenge gelernt werden. Man spricht hier von unüberwachtem Lernen oder ***unsupervised learning***. Clustering dient damit dem Auffinden unbekannter Inhalte im Sinne einer Hypothesenbildung [Chu+03]. Die Unterscheidung hinsichtlich des Lernverfahrens ist nicht streng, da auch Clusterverfahren danach beurteilt werden können, ob sie die erwarteten Ergebnisse bringen und entsprechend (manuell) angepasst werden.

Inwiefern sich mit Hilfe von Clusterverfahren komplexe Sachverhalte, also beispielsweise Vorgänge, erschließen lassen, ist Gegenstand der Arbeit von Kolz [Kolz06]. Dabei wurde deutlich, dass einerseits die entstehenden Cluster der sachkundigen Interpretation bedürfen und andererseits eher grobe semantische Zusammenhänge

identifiziert werden konnten. Dies konnte auch durch Betrachtung von Sätzen oder Teilsätzen nicht verbessert werden.

Das Ziel, ähnliche Elemente zu Kategorien zusammen zu fassen, erfordert die Entwicklung eines geeigneten Ähnlichkeitsmaßes. Den diesbezüglichen Überlegungen kommt zentrale Bedeutung zu, da wie erwähnt das Retrieval als Suche des zur Anfrage ähnlichsten Dokuments betrachtet wird.

2.4.2.1 Abstand und Ähnlichkeit

Ziel der Kategorisierung ist es, Objekte so zu gruppieren, dass sie sich innerhalb einer Kategorie minimal, zu allen anderen Objekten maximal unterscheiden. Damit stellt sich die Frage nach einem geeigneten Kriterium zur Beurteilung des Abstands und der Ähnlichkeit zweier Objekte zueinander.

Die Ermittlung dieser Größen erfolgt mit Hilfe einer Abstandsfunktion *d()*, wobei inhaltlich Abstand und Ähnlichkeit als invers zueinander, beispielsweise gemäß Formel 10 sind. Grundsätzlich entspricht ein geringer Abstand einer großen Ähnlichkeit und umgekehrt, wobei die Ähnlichkeit im Allgemeinen auf das Intervall [0;1] normiert ist.

Formel 10:Ähnlichkeit und Abstand bei normiertem bzw. nicht normiertem d

$$s(x,y) = 1 - d(x,y) \text{ bzw. } s(x,y) = \frac{1}{1 + d(x,y)}$$

Für *d()* gelten in Anlehnung an [Fisc89, S.10f] folgende Bedingungen:

- *Der minimale Abstand ist der eines Objektes zu sich selbst:* $d(x,x) = d_0 = 0$
- *Der Abstand zwischen zwei Objekten ist größer gleich* d_0: $d(x,y) \geq d_0$
- *Die Abstandsfunktion ist symmetrisch:* $d(x,y) = d(y,x)$
- *Die Dreiecksungleichung gilt:* $d(x,y) + d(y,z) \geq d(x,z)$

Zur Betrachtung von Ähnlichkeit können sowohl numerische als auch strukturelle Merkmale herangezogen werden, wobei entsprechend der wortorientierten Herangehensweise des traditionellen IR statistisch-numerische Merkmale die größere Bedeutung haben während strukturelle Ähnlichkeit insbesondere bei semantischen Netzen bzw. Ontologien interessiert.

Betrachtet man die Ähnlichkeit einzelner Worte, so sind die Anzahl gleicher Zeichenfolgen der Länge N, sogenannte N-Grams, ein häufig genutztes Kriterium. Wegen der allgemein gehaltenen Definition der N-Grams lassen sich analog auch Folgen von N Worten damit bewerten. Beispielsweise hat ‚Fenster' mit ‚bestens' die beiden 3-Grams ‚ens' und ‚ste' gemeinsam. Insgesamt bestehen die beiden Worte aus 8 verschiedenen 3-Grams, womit sich eine Ähnlichkeit von ¼ bzw. ein Abstand von ¾ ergibt.

Ausgehend von der Hamming-Distanz, die durch die Anzahl unterschiedlicher Stellen zweier Tupel definiert ist, hat Levenshtein [Leve66] ein häufig verwendetes Distanzmaß auf Wortebene entwickelt. Dieses auch als ***edit-distance*** bezeichnete Maß gibt die Kosten der Überführung eines Wortes in ein anderes wieder.

Um Längenunterschiede und klangliche Ähnlichkeiten zu berücksichtigen, werden unterschiedliche Kosten angesetzt. Im obigen Beispiel beträgt die Hamming-Distanz 6 unterschiedliche von insgesamt 7 möglichen Stellen, während die Levenshtein-Distanz bei Zugrundelegung von Kosten gemäß Tabelle 6 2,7 beträgt.

Tabelle 6: Änderungskosten auf Wortebene nach Levenshtein

‚Fenster'	Vorgang	Kosten	‚bestens'
F	ändern stimmlos	0,3	b
e	-	0,0	e
n	löschen	0,8	
s	-	0,0	s
t	-	0,0	t
e	-	0,0	e
r	ändern stimmhaft	0,6	n
	einfügen	1,0	s
	Gesamt:	2,7	

Aus Abschnitt 2.3.2.5 folgt, dass zwei Graphen genau dann gleich sind, wenn ihre Knoten- und Kantenmengen, eventuell unter Berücksichtigung des Richtungssinns, identisch sind. Von praktischem Interesse ist jedoch eher die als Isomorphie bezeichnete strukturelle Übereinstimmung zweier Graphen. Damit bietet der notwendige Aufwand zur Überführung eines Graphen in einen anderen eine Ausgangsbasis zur Ähnlichkeitsbestimmung. Oft stellt das Einfügen einer zusätzlichen oder das Ändern einer bestehenden Kante (z.B. Abbildung 16 links) einen geringeren Aufwand dar als das Hinzufügen eines Knotens. Daher ist eine Gewichtung der nötigen Schritte sinnvoll.

Tabelle 7: Ermittlung des Abstands zweier Graphen

Aktion	Kosten
Löschen Knoten B	1,0
Löschen Kante {B,C}	0,8
Hinzufügen Knoten F	1,0
Ändern Kante {A,B} in {A,F}	0,5
Gesamtkosten	3,3

Um beispielsweise in Abbildung 16 (S.63) den Graph a in Graph b zu überführen sind in Analogie zur ***edit-distance*** die in Tabelle 7 dargestellten Aktionen nötig, wobei der ermittelte Aufwand eventuell noch zu normieren ist, um die Größe der Graphen mit einzubeziehen. In Abhängigkeit von der Topologie des Graphen sowie der konkreten Anwendung spielen für die Betrachtung der Ähnlichkeit weitere Merkmale die Höhe bei Bäumen oder der minimale Weg zwischen zwei Knoten eine Rolle.

Nachdem der Abstand bzw. die Ähnlichkeit einzelner Objekte geklärt ist, stellt sich nun die Frage, wie diese Werte bei Mengen von Objekten zu verstehen sind. Dabei hat man prinzipiell die Möglichkeit, zwischen einem und alle Objekte zu berücksichtigen. Dies soll im Folgenden am Beispiel des Clustering dargestellt werden, wobei aus Gründen der Nachvollziehbarkeit nur der Fall betrachtet wird, ein Element einer von mehreren in Frage kommenden Gruppen zuzuweisen.

2.4.2.2 Abstand bei Clustern

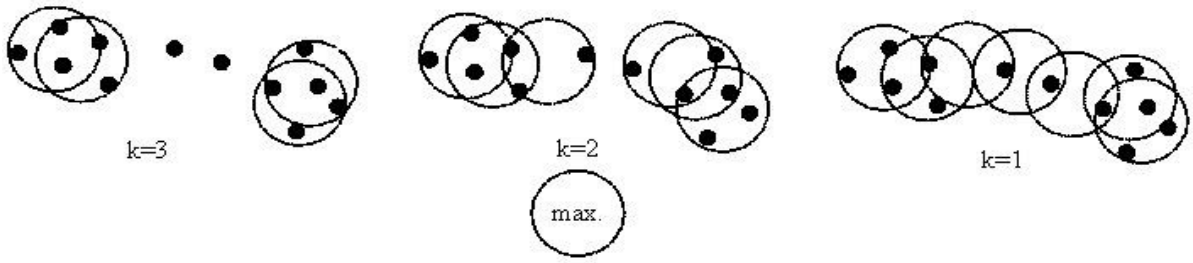
Beim ***single linkage*** wird ein Objekt einem Cluster zugewiesen, wenn es in diesem Cluster wenigstens ein Objekt in hinreichend geringer Distanz gibt. Die anderen Objekte spielen keine Rolle. Dadurch entstehen tendenziell ‚langgestreckte' Cluster und das Ergebnis ist von der Reihenfolge der betrachteten Objekte abhängig .

Fordert man, dass ein neu aufzunehmendes Objekt zu keinem der bisher im Cluster befindlichen eine Distanz größer d hat, spricht man von ***complete linkage***. Da innerhalb eines Clusters zwei Objekte einen maximalen Abstand von d zueinander haben, werden sich tendenziell ‚runde' Cluster ergeben. Da stets alle bereits zum Cluster gehörenden Objekte zu untersuchen sind, ist hier der Aufwand sehr groß.

Einen Kompromiss aus beiden Kriterien ist die Forderung, dass es innerhalb eines Clusters keinen Punkt geben darf, der nicht wenigstens k Nachbarn innerhalb der maximalen Distanz hat. Dieses Verfahren wird in der Literatur als ***k-nearest-neighbours*** bezeichnet. Sehr häufig wird k = 3 verwendet. Zu Details bezüglich dieser drei Kriterien sei auf die entsprechende Literatur wie [Ever+01, S.62ff] verwiesen.

Die Auswirkungen der unterschiedlichen Kriterien zeigt Abbildung 18, wobei der maximale Abstand konstant gehalten und durch einen Kreis mit Durchmesser *max* veranschaulicht wurde. Für abnehmende k bis ***single linkage*** erkennt man die Tendenz zur Bildung langgestreckter Cluster. Elemente außerhalb der Cluster werden als isolierte Punkte bezeichnet.

Abbildung 18: Auswirkung unterschiedlicher Clusterkriterien



Quelle: eigene Erstellung

Insbesondere bei natürlichsprachigen Texten besteht oft Ungewissheit, welcher Kategorie ein Dokument zuzuordnen ist. In diesen Fällen ordnet man ein Dokument nicht nur einem Cluster zu, sondern nutzt das jeweilige Ähnlichkeitsmaß im Sinne einer Zugehörigkeitsstärke. Damit ergeben sich Cluster mit unscharfen Grenzen, sogenannte ***Fuzzy-Clusters***. Diese Vorgehensweise ist auch bei Klassifikationen denkbar, wobei hier die Summe der Abweichungen von den Klassenkriterien (vgl. S. 67) als Abstand zu betrachten wäre.

Die Beurteilung der Qualität der Kategorisierung geschieht ähnlich dem Vorgehen nach Definition III und IV auf Seite 27: Eine Gruppe von Experten ordnet Dokumenten eines Referenzkorpus bestimmte Klassen zu bzw. fasst diese Dokumente zu Clustern zusammen. Je mehr Dokumente ein zu testender Algorithmus richtig der vorgegebenen Kategorie zuordnet, desto höhere Qualität wird dem Algorithmus unterstellt.

Orientieren sich die Algorithmen bei der Bildung der Cluster an einer Top-Down oder Bottom-up Vorgehensweise, werden diese als hierarchische Algorithmen bezeichnet. Top-Down bedeutet, dass die Gesamtmenge als ein einziger Cluster aufgefasst wird, der im Lauf der Verarbeitung in weitere Cluster gegliedert wird. Bei Bottom-Up werden alle Objekte als separate Cluster mit jeweils nur einem Element betrachtet und dann zu größeren Clustern zusammengefasst. Entsprechend werden diese Verfahren als diversive bzw. agglomerative Verfahren bezeichnet [Ever+01, S.55]. Im Fall der Klassifikation lassen sich Hierarchien durch die vorab definierte Hierarchie der Klassen realisieren.

Bei der Repräsentation von Texten in Form des Vektorraummodells erfolgt eine Abbildung von Termen auf Komponenten eines Vektors. Aus Sicht der Kategorisierung hat dies den Vorteil, dass alle Terme gleichberechtigt und gleichartig sind. Dadurch wird es möglich, die Kategorisierung auf eine Vielzahl gleichartiger, einfach aufgebauter Verarbeitungseinheiten zu verteilen.

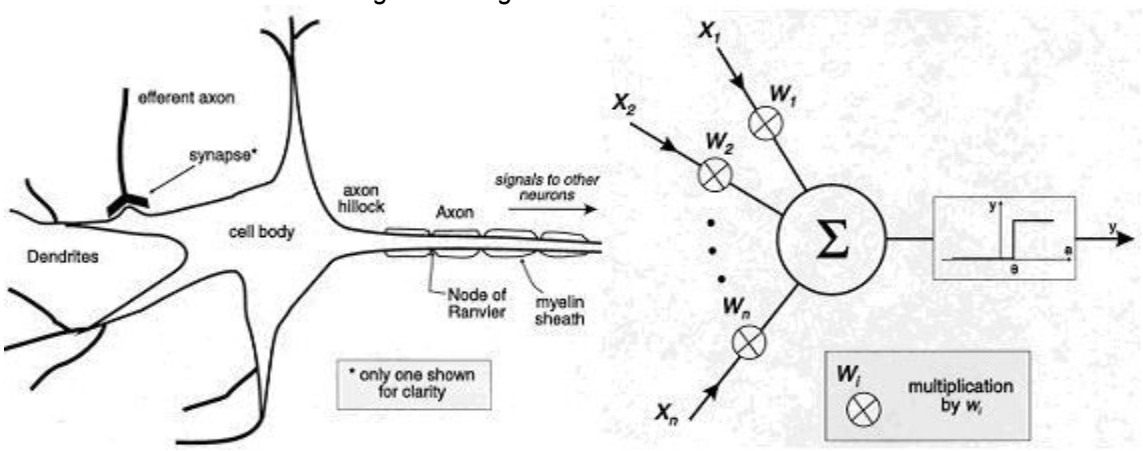
2.4.3 Konnektionistische Verfahren

Da das Gesamtergebnis aus dem Zusammenwirken vieler Einzelergebnisse entsteht, werden darauf aufbauende Verfahren als ***konnektionistisch*** bezeichnet. Die Repräsentation und Bearbeitung von Informationen erfolgt auf subsymbolischer Ebene, weswegen Rummelhart et al. dies auch als Mikrolevel der Modellierung bezeichnen [RuSm+87, S.7ff]. Einer der wichtigsten Vertreter sind die im Folgenden dargestellten künstlichen neuronalen Netze, wobei an dieser Stelle nur ein Überblick gegeben werden kann, wie er zum grundlegenden Verständnis des Prinzips notwendig ist.

2.4.3.1 Aufbau und Funktionsweise

Wie Abbildung 19 zeigt, orientiert sich der Aufbau der Verarbeitungseinheiten in künstlichen neuronalen Netzen stark an ihrem natürlichen Vorbild. Daher auch die Bezeichnung als ***Neuronen***. Wie diese beeinflussen sich nacheinander gelagerte Einheiten in Richtung der in einen Knoten eingehenden Kanten, wobei der Grad der gegenseitigen Beeinflussung im Modell durch die Gewichte w_i kontrolliert wird.

Abbildung 19: Biologisches und künstliches Neuron



Quelle: nach [Gurn97], S.3f/Kap.1

Je nach Gesamtstärke der Beeinflussung wird eine mehr oder weniger starke Aktivierung hervorgerufen. Ist die Aktivierung größer als ein Schwellwert θ, wird ein Reiz, typischer Weise digital als 0 oder 1 bzw. -1 und +1, ausgegeben. Um auf eine bestimmte Eingabe hin einen Reiz auszulösen ist eine Anpassung der Eingangsgewichte nötig, die dem Lernvorgang des natürlichen Vorbilds entspricht.

2.4.3.2 Anwendung zur Klassifikation

Aus technischer Sicht realisiert ein Neuron durch diesen Aufbau eine Abbildung eines n-dimensionalen Eingangsvektors auf die Menge {0,1} bzw. {-1,+1}, was der Klassifikation der Lage eines Punktes bezüglich einer Ebene des R^{n-1} entspricht.

Der Lernvorgang folgt dem Prinzip der Fehlerminimierung, wobei der Fehler hier Abweichung zwischen tatsächlicher y_{ist} und gewünschter Reaktion y_{soll} des Neurons ist. Damit ist klar, dass es sich um ein kontrolliertes Lernverfahren handelt, welches zur Klassifikation genutzt werden kann. Ein Lernfaktor α steuert, in welchem Maß ein Fehler zur Änderung der Gewichte führt. Formel 11 stellt das als Delta- oder auch Widrow-Hoff-Regel bezeichnete Lernverfahren dar.

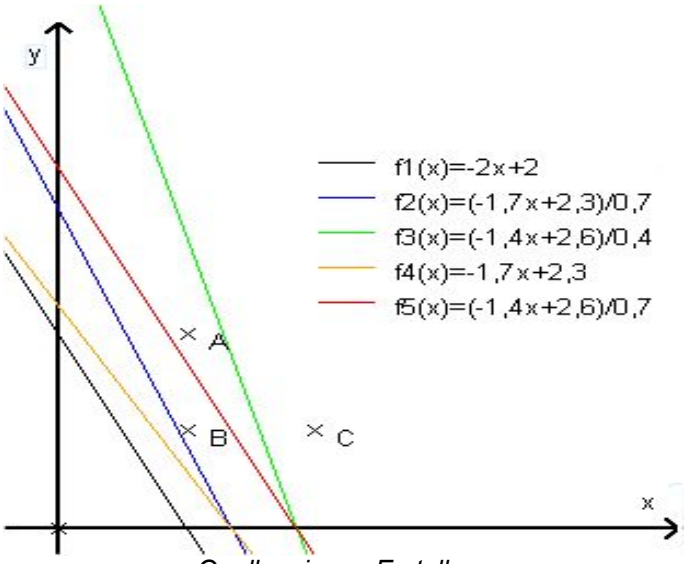
Formel 11: Lernregel nach [WiHo60]

$$w_i^{neu} = w_i^{alt} + x_i * (y_{soll} - y_{ist}) * \alpha$$

2.4.3 Konnektionistische Verfahren

Nach dem Lernen soll der Gesamtfehler für alle auftretenden Eingangsdaten, vor allem auch der nicht beim Training verwendeten, minimal sein [WrMo+03, S.523]. Abbildung 20 illustriert einen Lernvorgang am Beispiel der Trennung zwischen Punkt B und den Punkten A und C. Die Ebenen f_i sind zeitlich nacheinander aus f_1 hervorgegangen, wobei diese der Belegung $(\theta;w_1;w_2) = (2;2;1)$ entspricht. Als trennende Hyperebene hat das Neuron f_5, also $(\theta;w_1;w_2) = (2{,}6;1{,}4;0{,}7)$ gelernt.

Abbildung 20: Entwicklung der Trennebenen beim Perceptronlernen



Quelle: eigene Erstellung

Durch die Anordnung vieler Neuronen in einem Netzwerk verringert sich der Einfluss eines einzelnen Elements auf das Klassifikationsergebnis. Dies bewirkt einerseits eine Generalisierung der Muster, so dass auch ähnliche Eingangsvektoren korrekt klassifiziert werden. Andererseits beeinträchtigen Fehler wie der Ausfall eines Neurons oder ungeeignete Trainingsdaten nicht sofort das Gesamtergebnis. Man spricht diesbezüglich von ***graceful degradation***.

Mit der Erweiterung zu einem Netzwerk mehrerer Lagen von Neuronen geht ein weiterer Vorteil einher: die Klassifikationsergebnisse der einzelnen Einheiten lassen sich kombinieren, weswegen nicht mehr nur durch eine einzelne Hyperebene trennbare, also linear separable Merkmalsräume untersucht werden können. Auch aus Abbildung 20 geht bereits hervor, dass zur Trennung durch mehrere Hyperebenen mehrere Neuronen benötigt werden, deren Ergebnisse anschließend zusammenzu-

fassen sind. Zu weiterführenden Details sei auf die entsprechende Literatur wie beispielsweise [LäCl04] verwiesen.

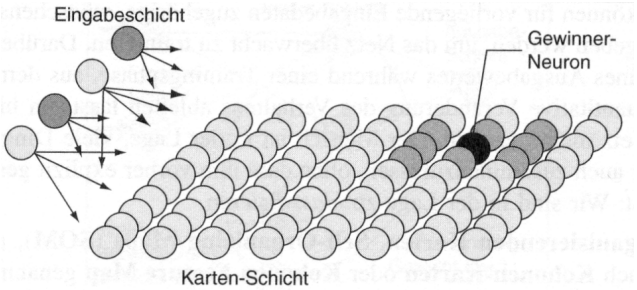
Die Schwierigkeit des praktischen Einsatzes eines neuronalen Netzes zur Klassifikation liegen vor allem in der Modellierung der Eingabe [Müll05]. Insbesondere bei der Anwendung auf Textdaten sind Informationen dazu nötig, welche Zeichenfolgen bzw. Textsegmente zu klassifizieren sind.

Besteht beispielsweise die Aufgabe, innerhalb eines Textes auftretende Bezeichnungen für Unternehmen zu identifizieren, so ist es sinnvoll, die in Frage kommenden Textstellen vorab zu kennzeichnen. Dadurch wird die durch die morphologische Vielfalt hervorgerufene Merkmalsmenge verallgemeinert und damit reduziert. Diese latente Klassifikation erfolgt beispielsweise durch einen Part-of-Speech-Tagger [Müll05, S.39f].

2.4.3.3 Anwendung zum Clustering

Auch für unüberwachte Lernverfahren und damit zum Clustering lassen sich neuronale Netze einsetzen. Bereits um 1970 entwickelte Kohonen dazu das Prinzip selbstorganisierender Karten [Koho01], deren Architektur Abbildung 21 zeigt.

Abbildung 21: Architektur einer Self Organizing Map



Quelle:[LäCl04, S.228]

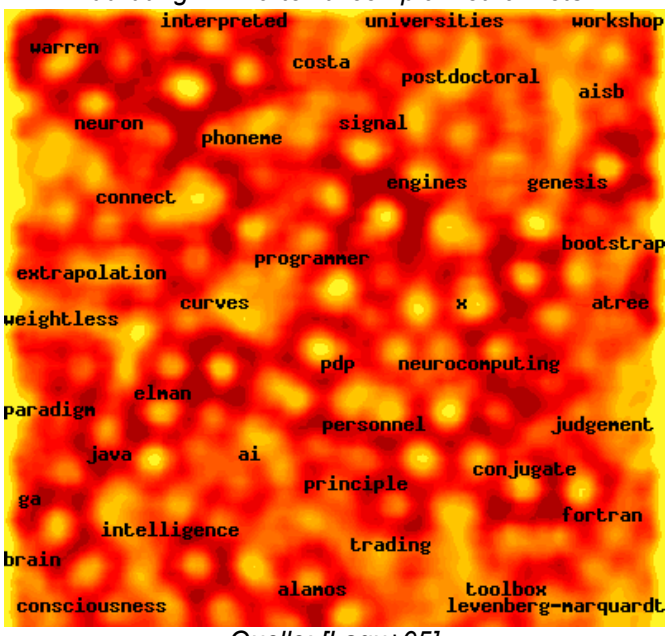
Kennzeichnend für diese als ***Self Organizing Map*** bezeichneten Netze ist eine Eingabeschicht, die dafür sorgt, dass jedes Neuron alle Signale der Eingabe erhält. Außerdem sind alle Neuronen untereinander verbunden. Zu Beginn des Lernvorgangs wird das Netz mit zufälligen Verbindungsgewichten initialisiert.

2.4.3 Konnektionistische Verfahren

Nach Anlegen einer Eingabe wird das am stärksten reagierende Neuron ermittelt. Innerhalb einer definierten, im Lauf des Trainings abnehmenden Umgebung um dieses Neuron wird gelernt indem die Gewichte aller Eingangskanten, die zum Reiz beigetragen haben, erhöht werden. Auch hier kommt wieder ein Lernfaktor zur Steuerung zum Einsatz.

In einer definierten, im Lauf des Trainings abnehmenden Nachbarschaft dieses Neurons wird ebenfalls gelernt, wobei die Stärke der Gewichtsänderung nach außen hin abnimmt. Im Ergebnis entsteht eine Karte, in der Muster großer Ähnlichkeit in einem Cluster liegen. Die Gegebenheiten des Merkmalsraums werden letztlich mit Hilfe der Gewichte auf die logische Anordnung der Neuronen abgebildet.

Abbildung 22: Karte zu 'comp.ai.neural-nets'



Quelle: [Lagu+05]

Abbildung 22 zeigt ein Self Organizing Map, welches zum Clustering von Newsgroup-Dokumenten benutzt wurde [Lagu00]. Die farbliche Darstellung entspricht der Aktivierung der Neuronen auf Grund der anliegenden Muster. Die Clusterzentren wurden mit einem statistisch ermittelten Signalwort aus den jeweiligen Beiträgen versehen.

Mehrlagige neuronale Netze zur Klassifikation nichtlinear separabler Merkmalsräume sowie selbstorganisierende Netze weisen auf Grund ihres Aufbaus einen Anlass zur Kritik auf: der Lernprozess, aber insbesondere die daraus hervorgehende Kategorisierung ist analytisch kaum nachvollziehbar [RuNo04, S.910].

2.4.3.4 Stützvektormethode

Die Stützvektormethode, oft als ***Support Vector Machine*** bezeichnet, ist ein statistisches Verfahren zur maschinellen Klassifikation. Es wurde um 1990 von Vapnik [Vapn95] entwickelt und trennt Merkmalsräume mit Hilfe von Hyperebenen. Wie beim neuronalen Netz wird auch hier mit 'Hilfe des Vorzeichens'[17] klassifiziert [WrMo+03, S.539]. Um auch nicht linear separable Merkmalsräume trennen zu können erfolgt bei diesem Ansatz eine Abbildung in Räume höherer Dimension. Erst dann erfolgt die lineare Trennung durch Konstruktion entsprechenden Hyperebene [Vapn95, S.133ff]. Vorgehen und Wirkungsweise skizziert Abbildung 23.

Abbildung 23: Stützvektormethode im linearen und nichtlinearen Fall

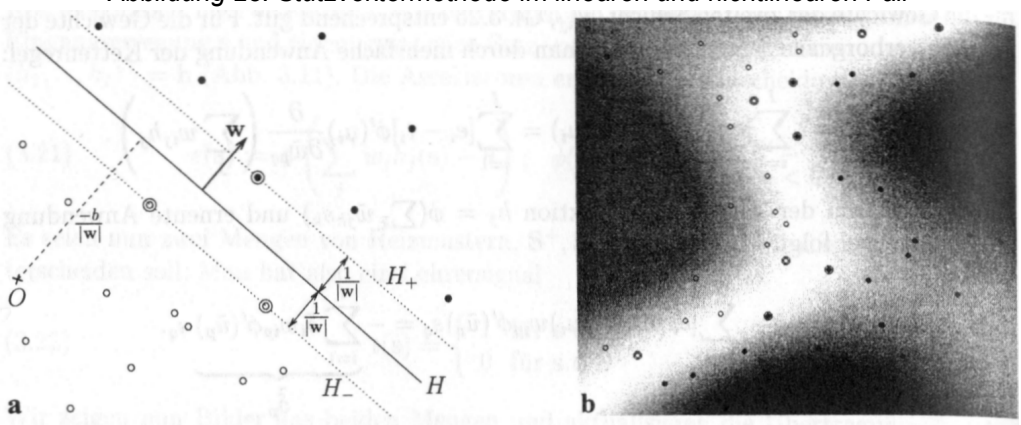


Abbildung 3.16: **a.** Prinzip der (linearen) SVM: die Supportvektoren (durch Kreise makiert) werden so bestimmt, dass die durch diese festgelegten Ebenen H_+ und H_- maximalen Abstand voneinenader haben. **b.** gekrümmte Trennfläche einer nichtlinearen SVM mit RBF-Kernel (3.35); erstellt mit dem JAVA-Applet von [Burges *et al.*, 1996].

Quelle: [MaHü+03, S.98]

Die Konstruktion dieser Hyperebene im transformierten Merkmalsraum ist sehr aufwändig. Boser et al. [Bose+92] haben jedoch gezeigt, dass man nicht diese Abbildung ermitteln muss, sondern das Skalarprodukt der abgebildeten Punkte genügt. Dies kann mit sogenannten Kernfunktionen sehr effizient ermittelt werden.

17 Ein in Richtung des Normalenvektors der Hyperebene liegender Punkt erzeugt ein positives Vorzeichen.

2.4.3 Konnektionistische Verfahren

Ziel künstlicher neuronaler Netze ist die Minimierung des Klassifikationsfehlers. Entsprechend wird der Lernvorgang beendet, sobald der Fehler hinreichend klein ist. Die Lage der trennenden Hyperebene im Raum spielt dabei keine Rolle. Bei der Stützvektormethode hingegen wird die Hyperebene so konstruiert, dass sie zu den nächsten Vektoren, den sogenannten Stützvektoren, maximalen Abstand aufweist.

Der Lernvorgang besteht daraus, die Parameter der trennenden Hyperebene zu bestimmen, was der Lösung eines quadratischen Optimierungsproblems entspricht [WrMo+03, S.541]. Desto mehr Trainingsdaten in die Optimierung eingehen, desto sicherer lässt sich die Lage der Hyperebene bestimmen. Da dies jedoch schnell zu Kapazitätsproblemen führt, wird die Trainingsmenge aufgeteilt und jeweils die Optimierung durchgeführt [MaHü+03, S.99]. An Hand der Vergleichsmenge wird die Güte der Klassifikation geprüft und gegebenenfalls eine neue Trainingsmenge ausgewählt [BeLi99, S.157f].

Bei geeigneter Modellierung lassen sich Vektoren mit mehreren tausend Dimensionen zuverlässig bearbeiten [May+03], was jedoch mit einem hohen Trainingsaufwand einher geht. Zur Anwendung in der Informationsextraktion, auf die in Abschnitt 2.5 genauer eingegangen wird, werden typischer Weise etwa 15 Merkmale herangezogen, so dass die Stützvektormethode sehr oft Anwendung findet und oft mit dem gesamten Gebiet des ***Machine Learning*** gleich gesetzt wird [IsKa02, Li+04]. Allerdings sind sehr große Mengen an Trainingsdaten nötig, damit die unterschiedlichen Klassen zuverlässig getrennt werden können.

2.4.4 Kategorisierung im Information Retrieval

Der Vorteil der Kategorisierung von Dokumenten, also deren thematischer Einordnung, liegt im Information Retrieval in erster Linie in der Eingrenzung des Suchraums. Lässt sich die Anfrage ‚Baum Aufbau' einem Thema zuordnen, müssen nur Dokumente eben dieses Themas weiter betrachtet werden. Dies führt zu einer Verringerung der Antwortzeit, zumal die Kategorisierung einmalig und vorab erfolgen kann. Außerdem werden Dokumentmengen dadurch strukturiert und somit für den Anwender leichter erschließbar. Ein Anwendungsfall ist die automatische Zuordnung von Emails in verschiedene Postfächer in Abhängigkeit vom Inhalt der Nachrichten.

Unscharfe Clustergrenzen eignen sich zur Auflösung von Ambiguitäten. Beträfe obige Anfrage zu 60% den Bereich Biologie und zu 40% die Graphentheorie, wäre es sinnvoll, die entsprechenden Anteile der Antwortmenge aus den beiden Klassen zu bilden. Damit erlaubt man dem Anwender, seine Anfrage im Nachhinein thematisch einzugrenzen. Eine solche Top-Down-Vorgehensweise ist beim Retrieval sinnvoll, wenn eine Eingrenzung auf relevante Dokumente nur schrittweise möglich ist. Dieses Vorgehen entspricht der Bildung unterschiedlicher Abstraktionsstufen.

Neben der Auszeichnung der Dokumente durch Metadaten können auch Ontologien genutzt werden, um die Clusterqualität zu verbessern. Dabei sind verschiedene Vorgehensweisen denkbar. Wird das Vektorraummodell zu Grunde gelegt, kommt eine Erweiterung des Dokumentenvektors um die bezeichneten Begriffe in Frage. Damit wird die Repräsentation in einen Term- und einen Begriffsteil gegliedert. Für den Fall der Sprachontologie ***WordNet*** [Mill+03] haben Hotho et al gezeigt, dass sich auf diese Art und Weise die Qualität des Clustering steigern lässt [HoSt+03].

Führt man die Klassifikation von Texten bis auf Satz- oder Wortebene fort, kommt man zum sogenannten ***Tagging***. Ähnlich der oben erwähnten Auszeichnung durch Metadaten können damit auch Sätze oder Satzteile mit beschreibenden Merkmalen gekennzeichnet werden, was letztlich der Identifikation bzw. Extraktion von Begriffen entspricht. Diesem Aspekt widmet sich der folgende Abschnitt.

2.5 Informationsextraktion

Informationsextraktion bezeichnet Verfahren, deren Ziel die Gewinnung von Informationen aus natürlichsprachigen Texten ist. Damit steht nicht mehr die Frage ***wo*** etwas zu finden ist im Mittelpunkt sondern die Frage ***welche Fakten*** im Text erwähnt werden. Diese Fakten betreffen interessierende, vorab festgelegte Arten von Ereignissen, also Entitäten und Vorgänge [Gate05], mit dem Ziel, Aussagen darüber zu formulieren. Innerhalb der Informationsextraktion beschäftigt sich ***Named Entity Recognition*** mit der Identifikation von durch Eigennamen bezeichneten Entitäten. Dies sind insbesondere Namen von Personen, Orten und Organisationen [CoNLL03], wobei letztere gelegentlich in Unternehmen und sonstige Organisationen untergliedert werden [McCa05].

Auf die extrahierten Informationen kann gezielt zugegriffen werden, beispielsweise zur automatischen Generierung von Inhaltsangaben. Die Informationsextraktion stellt damit eine ganz wesentliche Säule im Information Retrieval dar [FaNe98]. Insbesondere gilt dies unter dem Aspekt der Erschließung von Zusammenhängen bei Nutzung von Ontologien.

2.5.1 Problemstellung

Die grundlegende Schwierigkeit besteht darin, dass ein Sachverhalt auf sehr unterschiedliche Arten zum Ausdruck gebracht werden kann. Insbesondere Vorgänge werden oft durch Kombination mehrerer Worte wie ‚von ... auf' oder ‚zogen ... zurück' dargestellt. Die Vorgehensweise, Wortlisten sowie grammatische oder Assoziationsregeln zu erstellen, ist auf Grund dieser Vielfalt kaum anwendbar. Weiterhin werden viele Sachverhalte wie durch Präpositionen wie ‚von', ‚auf', ‚aus' zum Ausdruck gebracht, die jedoch bei Nutzung von Stoppwortlisten meist eliminiert werden.

Von Interesse sind also Verfahren, die ohne oder mit nur minimalen, vorab definierten Wortlisten bzw. Regelmengen arbeiten [Quas+02] und so den manuellen Aufwand gering halten. Die Art und Weise, wie Dinge zum Ausdruck gebracht werden, ist nicht statisch. Daher dürfen maschinelle Verfahren zur Erschließung von Sachverhalten nicht durch feste Vorgaben beispielsweise hinsichtlich des zu berücksichtigenden Vokabulars eingeschränkt sein.

Auf Grund der großen Bedeutung der Informationsextraktion als Voraussetzung eines tatsächlich auf Informationen zielenden Retrievals soll im Folgenden genauer auf die dabei notwendigen Schritte eingegangen werden.

2.5.2 Vorgehensweise

Die Extraktion von Informationen gliedert sich in fünf, oft eng ineinander greifende Teilaufgaben [McCa05]. Die ***Segmentierung*** bestimmt, welche Worte in die zu extrahierende Bezeichnung einfließen. Dies ist vor allem kritisch, wenn mehrere, morphologisch nicht zu unterscheidende Worte in Frage kommen. Im Rahmen der ***Klassifikation*** wird dieser Zeichenkette eine Bedeutung in Form eines Namens zugeordnet. Da für manche Klassen wie beispielsweise Ortsnamen nur bestimmte

Morphologien möglich sind, hängen hier Klassifikation und Segmentierung besonders eng zusammen.

Unter ***Assoziation*** versteht man die Extraktion von Beziehungen zwischen Informationseinheiten, die über mehrere Textabschnitte verteilt sind. Treten gleichartige Informationen wie Datums- oder Längenangaben in unterschiedlichen Notationen auf, so sind diese durch ***Normalisierung*** zu vereinheitlichen. Im Rahmen der ***Fehlerbereinigung*** werden Ambiguitäten aufgelöst und Fehlklassifikationen beseitigt.

Die Ergebnisse dieser Teilaufgaben äußern sich in Extraktionsregeln und beziehen sich wiederum auf interne und externe Hinweise. ***Interne Hinweise*** bezeichnen Merkmale innerhalb der Bezeichnung einer Entität. Ein Beispiel dafür ist die Rechtsform nach dem kennzeichnenden Namen einer Firma [Röss04]. Im einfachsten Fall existiert eine Liste aller denkbaren Merkmale. Wird ein entsprechender Eintrag im Text gefunden, kann man mit einer hohen Wahrscheinlichkeit davon ausgehen, dass im Beispiel unmittelbar davor der Name eines Unternehmens zu finden ist.

Als ***externe Hinweise*** bezeichnet man entsprechend Merkmale außerhalb der eigentlichen Bezeichnung. Dies sind zum einen Strukturmerkmale wie die Position im Text und zum anderen Triggerworte oder -wortarten, die in der näheren Umgebung einer Klasse von Eigennamen typischerweise auftreten. Damit gehören Hilfsmittel wie Wörterbücher sowie POS-Tagger prinzipiell ebenfalls zu externen Hinweisen. Ansätze zur ***Named Entity Recognition*** umfassen Wortlisten, Grammatikregeln, statistische Verfahren sowie Kombinationen davon [RaWa97, MaRa05]. Allerdings verlieren Wortlisten im Lauf der Zeit ihre Gültigkeit und sind – wie Statistiken – Abhängig von der betrachteten Domäne.

Externe Hinweise finden beispielsweise bei der Beurteilung der Relevanz von Internetseiten durch Suchmaschinen wie Google™ Verwendung [Bage05]. So wird hier unter anderem davon ausgegangen, dass wesentliche Informationen innerhalb der ersten vier Worte von Überschriften zu finden sind und unwichtige Informationen eher am Ende von Dokumenten stehen.

2.5.3 Stand der Technik

Die meisten Forschungsarbeiten auf dem Gebiet Informationsextraktion beschäftigen sich mit englischsprachigen Texten und erreichen da Werte für Recall und Precision von 90%, abhängig vom Texttyp und den zu erkennenden Klassen. Viele Sprachen, wie beispielsweise das Deutsche weisen jedoch eine reichere Morphologie mit vielen Unregelmäßigkeiten und einer eher freien Wortfolge auf. Die syntaktische Analyse ist daher fehleranfälliger. Weiterhin beginnen hier im Gegensatz zum Englischen nicht nur Eigennamen mit einem Großbuchstaben. Vielmehr sind alle Substantive bzw. substantivierten Worte potenzielle Eigennamen und fallen nicht zwingend in eine der drei üblicherweise betrachteten Klassen. Die Ergebnisse für Recall und Precision sind daher deutlich geringer [CoNLL03].

Viele kommerzielle Anwendungen und Forschungsprojekte wie Amilcare [Cira+95], basieren auf dem GATE-Framework der University of Sheffield [UnSh05]. Dabei handelt es sich um ein konfigurierbares System, welches mit Hilfe eines maschinellen Lernverfahrens einen Klassifikator zur Identifikation von Eigennamen erstellt. Diese Ansätze gehören somit zu den überwachten Lernverfahren und benötigen umfangreiche, manuell erstellte Trainingsdaten, erreichen aber oft sehr gute Ergebnisse.

Collins und Singer [CoSi99] haben ein Verfahren für englische Korpora entwickelt, welches ohne manuelle Eingriffe arbeitet. Allerdings sind hier gezielt auf Grund von Hintergrundwissen gewählte Ausgangsbeispiele, sogenannte ***Seeds***, und manuell erstellte Regeln notwendig. Dieses Vorgehen ist in hohem Maße sprachabhängig, da Bezeichnungen wie ‚Mr.' respektive ‚Herr' in deutschen Texten äußerst selten vorkommen und damit als Hinweis auf einen Personennamen eher ungeeignet sind. Gleiches gilt beispielsweise für die im Englischen oft genutzte Rechtsform eines Unternehmens. Part-of-Speech-Tagging (siehe S.19) stellt auch hier einen essentiellen Schritt der Vorverarbeitung dar.

Quasthoff et al [Quas+02] erarbeiteten an deutschen Texten ein Verfahren, um Regeln zur Beschreibung von Eigennamen zu generieren. Auch hier werden einige wenige initiale, manuell definierte Extraktionsregeln sowie Beispiele genutzt. Im Ergebnis entstehen Regeln zur Identifikation von Personennamen auf Grund mor-

phologischer Beschreibungen der umgebenden Worte. Allerdings baut dieses Verfahren auf Ergebnisse eines vorherigen POS-Tagging auf, so dass an dieser Stelle eine Verlagerung des analytischen Aufwands auf vorgelagerte Stufen stattfindet. Rössler [Röss04] stellte einen Ansatz vor, der völlig auf vorgegebene Regeln verzichtet, konzentrierte sich dabei allerdings nur auf Personennamen.

Ein kommerzielles Produkt im Bereich Information Retrieval, welches ebenfalls eine Komponente zur Extraktion von Informationen beinhaltet ist *orenge*™ [Empo05]. Das Gesamtprodukt ist dem Bereich ***Case Based Reasoning*** zuzuordnen, wobei eine Teilkomponente auf Fallbeschreibungen in Textform ausgerichtet ist. Dem System zu Grunde liegt eine durch eine Ontologie strukturierte Wissensbasis, die mit Hilfe von Kollokationsanalysen Unterstützung bei der Anpassung an den jeweiligen Diskursbereich bietet.

Die textbasierte Komponente gliedert sich in einen wortorientierten Teil, der mit Hilfe vorgegebener Schlüsselworte die Begriffe der Ontologie beschreibt sowie einen musterorientierten Teil von Extraktionsregeln. Letzterer identifiziert Begriffsinstanzen anhand vorzugebender regulärer Ausdrücke. Diese müssen, genau wie die Liste der Schlüsselworte, vollständig manuell erstellt werden. Weiterhin werden nur die exakten Muster ohne Generalisierung betrachtet. Als Beispiel wird die International Standard Book Number (ISBN) angegeben, die üblicherweise das Format 9-99999-999-9[18] hat. Häufig treten Abweichungen von dieser Struktur in Form unterschiedlicher Ziffernblockgrößen oder einem ‚X' an letzter Stelle auf, so dass für jeden Sonderfall manuell ein eigenes Muster zu erstellen ist[19].

Praktisch alle musterorientierten Ansätze benötigen vorheriges POS-Tagging, wie es beim System RAPIER [Cali98] vorgestellt wurde. Dabei ist nicht nur die grammatische Art des fraglichen Wortes sondern auch die der vorausgehenden und nachfolgenden Worte eine wichtige Entscheidungsgrundlage für die Klassifikation. Das bedeutet jedoch auch, dass falsche Ergebnisse der syntaktischen Analyse zu falschen Ergebnissen in der Klassifikation führen. Insbesondere bei Sprachen mit vielen Unregelmäßigkeiten und aufwändiger Grammatik tritt dieser Punkt immer

18 9 steht für eine beliebige Ziffer.
19 z.B. ISBN 3-540-60670-X: Frühwirth/Abdennadher – Constraint-Programmierung, Springer 1997.

mehr in den Vordergrund [Lezi99, Schm94]. Daher ist diese Vorgehensweise nur bei Verwendung eines verlässlichen Part-of-Speech-Taggers sinnvoll.

Außerdem ist zu beachten, dass für viele Sprachen überhaupt keine POS-Tagger existieren und darauf aufbauende Verfahren somit nicht eingesetzt werden können. Daraus ergibt sich die Notwendigkeit einer sprachunabhängigen Vorgehensweise, die möglichst ohne manuellen Aufwand zur Erstellung von Regeln oder Trainingsdaten auskommt.

Betrachtet man die Extraktion von Informationen aus kontextorientierter Sicht (S. 9), handelt es sich bei den Entitäten um Begriffe und bei deren Bezeichnungen entsprechend um Begriffsinstanzen. Damit stellt die Informationsextraktion eine essenzielle Komponente begrifflicher Erschließung von Texten dar. Insbesondere ebnet diese begriffsorientierte Auffassung den Weg zur Erschließung abstrakter Begriffe wie ‚Geschäftsfeld' oder ‚Rolle', die über die ***Named Entity Recognition*** hinaus gehen [Lore05].

Entsprechend dieser Sichtweise auf den Begriff ergibt sich damit insgesamt die Überlegung, Kommunikationsmuster im Sinne von Begriffsverbänden maschinell zu erschließen. Dadurch kann die eingangs dargestellte Problematik der verschiedenen Ausdrucksmöglichkeiten zwar nicht vollständig eliminiert werden. Allerdings ist zu erwarten, dass innerhalb eines Diskursbereichs bestimmte Ausdrucksweisen typischer als andere sind. Bei gegebener Eingrenzung der Domäne sollte eine maschinelle Erschließung der Kommunikationsmuster und darauf aufbauend die Klassifikation der auftretenden Begriffsinstanzen also sehr effizient sein. Durch die Nutzung der tatsächlich auftretenden Kommunikationsmuster ist weiterhin zu erwarten, dass die zu Grunde liegende Sprache unerheblich für das Verfahren ist.

2.6 Fazit

2.6.1 Problembeschreibung

Derzeit lassen sich im Bereich des Information Retrieval zwei Hauptrichtungen identifizieren. Zum einen die klassische, termbasierte Vorgehensweise, die darauf abzielt, Verweise auf Texte zu finden, die für eine Anfrage relevant sind. Basis ist

hier die Bestimmung der Ähnlichkeit zwischen einer Anfrage und einem Dokument, wobei beide in Form von Vektoren aus gewichteten Termen dargestellt werden. Zur Identifikation und Gewichtung relevanter Terme dienen vor allem statistische Verfahren.

Der Vorteil dieser auf einem extensionalen Begriffsverständnis fundierten Herangehensweise ist die einfache Informationsgewinnung durch Selektion einzelner Terme. Der damit verbundene Nachteil ist das Fehlen jeglicher Zusammenhänge. Durch Kategorisierung von Termen wird versucht, Hinweise auf Zusammengehörigkeiten zu gewinnen.

Die zweite auszumachende Richtung adressiert diese Problematik des fehlenden semantischen Zusammenhangs. Wesentliches Merkmal ist daher die explizite Modellierung des untersuchten Diskursbereichs mit Hilfe ontologischer Modelle wie beispielsweise semantischen Netzen. Da deren maschinelle Erstellung mit Hilfe statistischer Verfahren noch immer experimenteller Natur ist [Maed+03, CiHo+05], ist ein hoher manueller Aufwand nötig.

Um die Vorteile der einfachen Informationsgewinnung beim Retrieval und des gezielten, schnellen Zugriffs zu verbinden, wurden Ansätze entwickelt, die das Vektorraummodell um begriffliche Merkmale erweitern [HoSt+03]. Das weiterhin bestehende Problem liegt jedoch darin, diese semantischen Merkmale maschinell zu gewinnen, sofern es sich nicht lediglich um Metainformationen wie Datums-, Quellen- oder Autorangaben handelt.

Durch die Modellierung erreicht man eine einheitliche Sprache von Informationsangebot und -nachfrage. Darüber hinaus einigen sich damit alle am Informationsprozess beteiligten Personen auf eine gemeinsame Begriffswelt. Diese Begriffe sind nicht nur Eigennamen sondern umfassen auch abstrakte Konzepte und Vorgänge. Damit einher geht eine Eingrenzung und Strukturierung des Diskursbereichs, so dass die Modellierung aus Sicht des Information Retrieval nützlich ist.

Allerdings ist anzumerken, dass mit dieser Modellierung stets der aktuelle Zustand eingefroren wird, was bei geändertem Informationsbedarf eine Änderung oder gar erneute Modellierung bedingt. Daraus folgt, dass vor allem stabile, empirische Begriffe Bestandteil des Modells sein sollten [Czap88, S.216f].

2.6.1 Problembeschreibung

Durch Ontologien modellierte Begriffssysteme können nur strukturierte, identifizierte Informationen erfassen. In einem Text treten jedoch im Allgemeinen nicht nur die definierten Begriffsbezeichnungen sondern eine Vielzahl weiterer Bezeichner auch anderer, nicht interessierender Begriffe auf. Damit stellt sich die Aufgabe, die in einem Text vorkommenden Bezeichner als Instanzen der definierten Begriffe zu identifizieren und zu extrahieren um so ein Matching zwischen modelliertem Ideal und niedergeschriebener Realität zu erreichen.

Ein Ansatz dazu ist die Verwendung von Sprachontologien wie ***WordNet*** [Mill+03] oder GermaNet [GerNe04]. Allerdings ist die Modellierung des gesamten Vokabulars einer Sprache einerseits sehr aufwändig und andererseits stets unvollständig. Außerdem geht damit eine Einschränkung auf eben diese modellierte Sprache einher. Ähnliches gilt für die Verwendung von Part-of-Speech-Taggern im Rahmen der Informationsextraktion, da auch dies die sprachlich neutrale Anwendbarkeit einschränkt.

Es stellt sich die Frage, ob die Ursache der beschriebenen Schwierigkeiten beider Herangehensweisen doch nicht auf technischer Seite liegen und daher auch nicht durch verbesserte Algorithmen zu beheben sind. Den dargestellten Techniken zu Grunde liegt das Verständnis vom Wesen eines Begriffs. Daher ist zu untersuchen, ob nicht ein anderes Begriffsverständnis als das üblicherweise angenommene Extensionale hilfreich sein kann.

2.6.2 Lösungsansatz

Die in Abschnitt 2.1.1.2 dargestellte Auffassung vom Begriff geht davon aus, dass sich Begriffe durch ihre Verwendung in der Sprache definieren. Ein wesentlicher Aspekt ist dabei das gemeinsame Auftreten der Begriffe innerhalb einer Äußerung.

Eine Voraussetzung effizienter Kommunikation ist, dass die übertragenen Aussagen verständlich, also interpretierbar sind. Aus diesem Grund gehorchen Sprachen als Kommunikationsmittel bestimmten Regelmäßigkeiten, die sich in ihrer Grammatik widerspiegeln. Für das Information Retrieval interessiert jedoch nicht die Grammatik in ihren Details, sondern die damit bewirkte Anordnung von Worten. Aus begrifflicher Sicht handelt es sich bei diesen Worten um Bezeichner von Begriffen, die auf Grund der Grammatik in Kommunikationsmustern angeordnet sind.

Die Bedeutung der Begriffe kann unter verschiedenen Aspekten betrachtet werden. Aus syntaktischer Sicht als Substantiv, Verb, etc., aus Sicht ihrer linguistischen Funktion als Subjekt, Prädikat, Objekt oder eben als Begriffe im Sinne von Vorstellungen von Sachverhalten. Vor diesem Hintergrund ist ein Text eine Anordnung von Begriffsbezeichnern, die gewissen Mustern genügt. Sind diese Muster bekannt, lässt sich die Bedeutung der darin auftretenden Bezeichner erschließen, wie das Beispiel auf Seite 9 zeigte. Der Bezeichner an sich, also die Zeichenfolge, ist austauschbar.

Im Rahmen der Modellierung wird festgelegt, welche Begriffe für die konkrete Anwendung interessant sind. Damit einher geht im Allgemeinen, dass man für diese interessierenden Begriffe auch Beispiele für Bezeichner nennen kann. An Hand dieser Beispiele sollten sich wiederum die Kommunikationsmuster identifizieren lassen. Damit eröffnet sich die Möglichkeit, ausgehend von einer kleinen Anzahl bekannter Bezeichnungen auf die Bedeutung weiterer Bezeichner zu schließen. So können inkrementell die verwendeten Kommunikationsmuster sowie weitere Bezeichner der interessierenden Begriffe erschlossen werden.

Aufbauend auf den im Grundlagenteil dargestellten Methoden und den daraus gefolgerten Ableitungen widmet sich der folgende Teil der Arbeit der Ausarbeitung dieses Ansatzes. Innerhalb der prototypischen Realisierung werden diese Darstellungen einer praktischen Prüfung unterzogen und die damit zu erreichenden Ergebnisse bewertet.

Das Hauptaugenmerk liegt dabei auf der Konzeption, während die Umsetzung neben der Bewertung des Verfahrens vor allem praktische Details, auftretende Probleme und deren Lösung demonstrieren soll. Außerdem bietet die Umsetzung einen ersten Anhaltspunkt zur Bewertung der Potentiale des Ansatzes und zeigt Möglichkeiten für Verbesserungen auf. In die Konzeption sollen folgende, eng miteinander verknüpfte Rahmenbedingungen eingehen, die die Notwendigkeit und den Einfluss externer Eingriffe minimieren:

2.6.2 Lösungsansatz

- Sprachunabhängigkeit: Neben dem Verzicht auf syntaktische Vorarbeiten mit Hilfe eines POS-Taggers dürfen auch keine Wortlisten genutzt werden. Sprachunabhängigkeit in diesem Zusammenhang bedeutet, dass die Vorgehensweise an sich auf verschiedene Sprachen übertragbar ist.
- Flexibilität: Die manuelle Erstellung von Extraktionsregeln bedeutet, neben demhohen Erstellungsaufwand, dass die Anwendung auf einen bzw. wenige Diskursbereiche eingeschränkt wird. Auch die Neutralität bezüglich der zu Grunde liegenden Sprache ginge damit verloren.
- Nutzbarkeit: Desto mehr Informationen zur Einrichtung einer Anwendung nötig sind, desto geringer ist in der Summe der gewonnene Nutzen. Dementsprechend dürfen nur wenige, durch den sachkundigen Anwender leicht beizubringende Ausgangsinformationen zur Initialisierung nötig sein. Umfangreiche Trainingskorpora wie bei statistischen Verfahren des ***Machine Learning*** (vgl. S.38) verbieten sich daher.

3 Konzeption und Realisierung

Dieses Kapitel folgt einer Top-Down-Vorgehensweise zur Beschreibung der Konzeption und deren Realisierung. Anhand der im Grundlagenteil dargestellten Aspekte und Methoden wird zuerst umrissen, welche davon wie zur Realisierung eines begriffsbasierten Information Retrieval geeignet sind. Darauf wird die Konzeption aufgebaut, die auf Details eingeht und Anknüpfungspunkte für weitergehende Arbeiten aufzeigt. Die unterste Ebene ist die Implementierung, die sich den Aspekten der Programmierung widmet.

3.1 Ableitungen aus den Grundlagen

3.1.1 Eingrenzung

Ziel der prototypischen Umsetzung ist die Überprüfung der prinzipiellen Vorgehensweise hinsichtlich Funktionalität und Realisierbarkeit. Eine vollständige Umsetzung des Konzepts im Form eines System zum Information Retrieval würde die für diese Arbeit gegebenen Ressourcen übersteigen. Aus diesem Grund konzentriert sich dieser Teil auf die Kernaspekte und zeigt an den entsprechenden Stellen auf, wie eine notwendige Erweiterung bzw. Anpassung erfolgen muss.

Der Kernpunkt der Arbeit ist die Überlegung, dass Kommunikation über Muster von Begriffen funktioniert. Wir wissen, wie wir Sachverhalte ausdrücken müssen damit sie verstanden werden. Dieses Wissen erlaubt uns, auch unbekannte Sachverhalte richtig zu interpretieren. Dadurch verstehen wir Aussagen wie ‚sagte Ranchit Özman gestern in Al-Luhah', obwohl uns weder Person noch Ort bekannt sind[20].

Dieses Prinzip gilt auf Ebene von Einzelworten, wo es die Klassifikation unbekannter Einheiten (Entitäten) als Instanzen von Begriffen erlaubt, und widerspiegelt auf höherer Ebene den Zusammenhang zwischen Entitäten. Im Beispiel wird uns – je nach Kenntnis der Wortbedeutung – mitgeteilt, dass ein Zusammenhang zwischen einer Person und einem Ort besteht. Desto genauer die Kenntnis auf Wortebene ist, desto genauer erschließt sich der Zusammenhang zwischen den Worten.

20 Der an dieser Stelle mögliche Einwand, dass es sich nicht zwingend um eine Person bzw. einen Ort handelt zeigt, dass es tatsächlich funktioniert...

3.1.2 Begriffssicht

Wie im Fazit angedeutet soll die in Kapitel 2.1.1 als ‚kontextorientierte Sicht' dargestellte Begriffsauffassung zu Grunde gelegt werden. Bei dieser Sicht tritt das eigentliche Wort in den Hintergrund und Kommunikation wird als Zusammenspiel von Begriffen verstanden [Pole78, S.161]. Dieses Zusammenspiel folgt typischen Mustern um eine effektive und effiziente Kommunikation zu gewährleisten. Schon mit einer kleinen Anzahl bekannter Bezeichner für die zu betrachtenden Begriffe sollten sich diese Muster in natürlichsprachigen Texten identifizieren lassen. Mit Hilfe dieser Muster kann dann auf die Bedeutung weiterer Bezeichner geschlossen werden.

Um zu gewährleisten, dass nicht zufällige Konstellationen als gültige Muster betrachtet werden, muss die Textmenge so groß sein, dass die Muster innerhalb dieses Korpus mehrfach auftreten. Weiterhin ist eine Einschränkung auf eine Domäne und somit Fachsprache notwendig. Nur dann ist davon auszugehen, dass die interessierenden Begriffe in hinreichender Häufigkeit und in widerspruchsfreien Konstellationen auftreten.

Dieser Aspekt wird durch den zu betrachtenden Texttyp beeinflusst. Nachrichten beispielsweise dienen in erster Linie der Weitergabe von Informationen und versprechen daher eine hohe Informations- und damit Begriffsdichte. Bei dieser Textsorte ist weiterhin davon auszugehen, dass sich Kommunikationsmuster besonders leicht identifizieren lassen, wenn man Sprache als Kommunikation durch Muster von Begriffen versteht. Aus diesen Gründen sollen für die Realisierung Nachrichtenkorpora wie das Archiv von heise-online und Reuters verwendet werden.

Im Rahmen der prototypischen Umsetzung können nur wenige Begriffe in Betracht gezogen werden, was mit einem geringen Detaillierungsgrad einher geht. Es ist jedoch zu erwarten, dass sich Ansatzpunkte zur Gewinnung von Hyponymen sowie synonymen Bezeichnungen auf gleicher Abstraktionsebene ergeben werden. Weiterhin ist davon aus zu gehen, dass Fehlklassifikationen auf Grund sprachlicher Besonderheiten auftreten werden.

So ist beispielsweise zu erwarten, dass ‚Deutschland' als Firma klassifiziert wird, wenn die Verwendung in den Dokumenten dies nahe legt und es einen Begriff <STAAT> in der definierten Begriffswelt nicht gibt. Dieses Problem tritt jedoch auch bei manuell erstellten Regeln nach Collins und Singer [CoSi99] auf, wenn beispielsweise ‚Deutschland AG' im Text vorkommt. Der Grund dafür ist in der Pragmatik der Sprache zu sehen, die der beabsichtigten Wirkung einer Äußerung folgt.

3.1.3 Sprachverständnis

Aufbauend auf dem dargestellten Begriffsverständnis ist es nicht die Sprache an sich, die es zu verstehen gilt. Nur die sprachlichen Mittel der Kommunikation, die ihre Entsprechung in den Kommunikationsmustern haben, sind zu identifizieren. Zur Beschreibung dieser Muster genügen die Möglichkeiten einer Typ-3-Grammatik, konkret regulärer Ausdrücke.

Die thematische Einschränkung bringt in diesem Zusammenhang einen weiteren Vorteil, weil das damit verbundene Fachvokabular die Vielfalt reduziert. So wird man in Fachtexten aus dem Bereich Wirtschaft kaum auf Ausdrücke wie ‚sich artikulieren' oder ‚reden' stoßen, wenn auf Aussagen von Personen referenziert wird. Stattdessen findet man sehr oft Formulierungen mit ‚sagen' und ‚erklären'[21]. Diese Formulierungen geben auf Grund ihrer im Vokabular begründeten Häufigkeit und zeitlichen Stabilität einen externen Hinweis (siehe S.81) auf die Kommuikationsmuster bzw. die darin enthaltenen Begriffsbezeichner [Domi+01].

Zur Beschreibung externer Hinweise lassen sich reguläre Ausdrücke (siehe S.17) nutzen. Die wichtigsten, zum weiteren Verständnis nötigen Beschreibungen regulärer Ausdrücke zeigt Tabelle 8. Insbesondere hinsichtlich technischer Erweiterungen wie der Fähigkeit, Ergebnisse zu speichern, sei auf [Schmi02] verwiesen.

Ein solcher regulärer Ausdruck, der auf den Begriff ‚Aussage' hinweist, könnte lauten: , *(?:sag|erklär)[ten]+'* und würde alle Zeichenfolgen erkennen, die mit einem Leerzeichen gefolgt von der Zeichenfolge ‚sag' oder ‚erklär' beginnen und auf mindestens eines der Zeichen t, e und n oder eine beliebig lange Kombinationen daraus enden. Damit wären zwar neben den tatsächlich sinnvollen Worten auch

21 Im Korpus K2003 tritt knapp 4000 mal sag* und 1300 mal erklär* auf, jedoch nur 1 mal artikulier* und 140 mal rede*.

3.1.3 Sprachverständnis

Formen wie , sagnettenteen‘ abgedeckt. Da es hier aber nicht um eine Generierung sondern um eine Erkennung von Sprache geht, scheint es adäquat, sich auf das Nichtauftreten dieser Sonderfälle zu verlassen.

Tabelle 8: Syntax regulärer Ausdrücke im Überblick

Ausdruck		Beschreibung	
elementar		*Allgemeiner regulärer Ausdruck*	
	x		Einzelnes Zeichen 'x'
	Wort		Zeichenfolge 'Wort'
	[xyz]		Zeichenklasse: 'x', 'y' oder 'z'
	[a-z]		Zeichenklasse: 'a' bis 'z'
	[^x]		Alles, was nicht 'x' ist
	(Wort)		Zeichenfolge 'Wort' gruppieren und speichern
	(?:der|das)		Zeichenfolge 'der' oder 'das' gruppieren aber nicht speichern
modifizierend		*Modifiziert den vorhergehenden elementaren Ausdruck*	
	*		0 bis n mal
	+		1 bis n mal
	?		0 oder 1 mal
	{4}		Genau 4 mal
	{2,7}		2 bis 7 mal

Die eigentlichen Kommunikationsinhalte sind im zeitlichen Verlauf weniger stabil. Unter dem Gesichtspunkt der Entropie (S.40) ist der Informationsgehalt umso höher, je häufiger sie sich ändern. Die Mitteilung über den Namen des Bundeskanzlers ist nur informativ, wenn sie eine Änderung der bisherigen Kenntnis bedeutet. An dieser Stelle ist also lediglich eine generelle morphologische Beschreibung möglicher Inhalte anzugeben, was ebenfalls mit Hilfe einer Typ-3-Grammatik erfolgen kann. Inwieweit dies möglich ist, hängt von konkreten Anwendungsfall ab. So weisen insbesondere Bezeichnungen von Unternehmen eine sehr große Vielfalt auf, die kaum erschöpfend so dargestellt werden kann [Müll05, S.40]. Andererseits eignen sich solche spezifischen Besonderheiten als interne Hinweise (siehe S.81) zur Klassifikation der fraglichen Bezeichner.

Insgesamt ergibt sich eine zweiteilige Beschreibung der Kommunikationsmuster mit Hilfe von regulären Ausdrücken. Sie gliedern sich in einen stabilen, als Makromuster bezeichneten, umgebenden Teil sowie einen instabilen, informationstragenden Teil, der als Mikromuster bezeichnet wird. ‚Stabil‘ bedeutet in diesem Zusammen-

hang soviel wie ‚immer wieder auftretend' [Domi+01]. Dies soll zum Ausdruck bringen, dass der zu identifizierende Begriff bzw. seine unterschiedlichen Bezeichner regelmäßig in diesem stabilen Umfeld auftreten. Aus linguistischer Sicht ist diese Stabilität in etwa mit der Lexikalisierung von Worten vergleichbar [Lehm06]. Da beide Komponenten des Musters unmittelbar aus Sprachverständnis folgen, werden sie in den folgenden Abschnitten näher erläutert.

3.1.3.1 Mikromuster

Viele Begriffsbezeichnungen genügen aus morphologischer Sicht einem bestimmten Muster. Dies sei am Bezeichner für ‚Person' erläutert: Personennamen bestehen mindestens aus Vor- und Zunamen, die jeweils mit einem Großbuchstaben beginnen und durch ein Leerzeichen getrennt sind. Sowohl Vor- als auch Zunamen können zweigliedrig sein, wobei die Verbindung durch einen Bindestich erfolgt. Zwischen Vor- und Zuname können second initials oder Abstammungsbezeichnungen auftreten. Der reguläre Ausdruck dazu lautet:

[A-Z][a-z]+(-[A-Z][a-z]+)? ([A-Z]. |von (der)?)?[A-Z][a-z]+(-[A-Z][a-z]+)?

Je spezifischer eine derartige Beschreibung ausfällt, desto geringer ist die Wahrscheinlichkeit, dass sie auf zufällige Konstellationen passt. Entsprechend unwahrscheinlich ist es, dass die Beschreibung einer ISBN (siehe Beispiel S. 83) auf andere Zeichenkombination passt. Für derartige Instanzen wie auch Geldbeträge und Datumsangaben genügt also ein Mikromuster zur Extraktion.

Die Mehrzahl der in Frage kommenden Begriffe lässt sich jedoch deutlich einfacher, beispielsweise im Deutschen etwa als ‚Wort mit großem Anfangsbuchstaben' beschreiben, was sich in dem regulären Ausdruck ‚[A-Z][a-z]+' widerspiegelt.

3.1.3.2 Makromuster

Bei solchen wenig spezifischen Mikromustern von kommt dem Umfeld, in dem eine Begriffsinstanz auftritt eine noch größere Bedeutung zu. Dieses Umfeld wird im Folgenden als Makromuster bezeichnet. Erst das gesamte Kommunikationsmuster ist zur Identifikation einer solchen Begriffsinstanz geeignet. Califf bezeichnet dieses Umfeld als pre- bzw. postfiller, wobei hier als zusätzliche Merkmale wie insbesonde-

re die durch syntaktische Analysen ermittelten Wortarten betrachtet werden [Cali98].

Ein Makromuster beinhaltet demnach ein Mikromuster und lässt sich ebenfalls durch einen regulären Ausdruck beschreiben. Das Prinzip sei an folgendem Beispiel illustriert:

> *„... Über 256 GFlops soll der Cell in grafischen und visuellen Anwendungen leisten ... Ein bei IBM in East Fishkill hergestellter Prototyp hat Taktgeschwindigkeiten bis zu 4,6 GHz erreicht. Der Pilotbaustein wird in einem 90-Nanometer-Prozess produziert. Die 234 Millionen Transistoren haben dabei auf einem Plättchen von 221 Quadratmillimeter Platz. ...“*
>
> *[HON05b]*

Um aus diesem Textausschnitt den Bezeichner der Strukturgröße <STRUK> zu extrahieren genügt es nicht, mit Hilfe des Mikromusters nach zwei bis drei aufeinander folgenden Ziffern ([0-9]{2,3}) zu suchen. Das korrekte Ergebnis findet man nur, wenn zusätzlich die Nennung der Skalierung gefordert wird, womit sich das Makromuster ‚[0-9]{2,3}-Nanometer‘ und entsprechend das Kommunikationsmuster ‚<STRUK>-Nanometer‘ zur Weitergabe einer Strukturgröße ergibt.

Es leuchtet unmittelbar ein, dass mit der zunehmenden Vergrößerung und damit Spezialisierung der Muster, die Precision der Ergebnisse tendenziell zunimmt. Die zu erwartende Anzahl identifizierter Instanzen, der Recall, wird hingegen abnehmen. Um letzterem entgegen zu wirken, sind die Makromuster zu generalisieren. So wird obiges Muster mehr Instanzen des Begriffs ‚Strukturgröße‘ identifizieren, wenn es zu ‚<STRUK>-(Nano|Mikro)meter‘ generalisiert wird.

Wie in diesem Beispiel ist es wenig sinnvoll, lediglich die durch das Mikromuster beschriebenen Komponenten zu extrahieren. Die Angabe ‚90‘ allein ist zu wenig aussagekräftig. Daher bietet es sich hier an, die gesamte Zeichenfolge ‚90-Nanometer‘ zu extrahieren.

3.1.4 Lernen der Muster

Da die Kommunikationsmuster nicht manuell vorgegeben werden, ist ein Lernverfahren zu entwickeln. Als weitere Rahmenbedingung gilt, dass dieses Verfahren nicht auf große Trainingskorpora angewiesen sein darf. Aus technischer Sicht muss das Verfahren lernen, Zeichen- oder Wortfolgen als Bezeichnungen von Begriffen zu klassifizieren. Die Entdeckung neuer Begriffe ist nicht Gegenstand dieser Arbeit. Lediglich weitere Abstraktionsstufen innerhalb der gegebenen Begriffe sollen in Betracht kommen. Damit gestaltet sich die prinzipielle Vorgehensweise wie folgt:

Für jeden Begriff wird eine kleine Menge von Begriffsinstanzen, also typischen Bezeichnern dieser Begriffe, vorgegeben. Da diese Beispiele den Ausgangspunkt des Lernverfahrens bilden, wird diese Menge als ***Seed*** bezeichnet. Diese dienen der Auszeichnung einer Sprachprobe – in diesem Fall eines Korpus – mit den jeweiligen Begriffen. Der ausgezeichnete Korpus wird nach Mustern durchsucht, die den prinzipiellen Aufbau ‚*Freitext <BEGRIFF> Freitext*' haben.

Besonders interessant sind Muster mit mehreren Begriffen, die zueinander in Beziehung stehen. Ein Beispiel dafür sind Personen, Unternehmen und die Funktionen von Personen in Unternehmen. Dies führt zu Kommunikationsmustern mit mehreren Begriffen. Aus kontextorientierter Sicht hat ein solches mehrstelliges Muster eine größere Aussagekraft, je mehr Begriffe daran beteiligt sind.

Zur Anwendung der Muster wird die Begriffsbezeichnung durch das entsprechende Mikromuster ersetzt. Das resultierende Makromuster wird auf den Korpus in einer Volltextsuche angewandt und die dem Mikromuster genügenden Zeichenfolgen extrahiert. Die gewonnenen Fakten sowie erfolgreich angewandte Muster werden in einer Wissensbasis abgelegt und dienen in einem nachfolgenden Schritt des Lernzyklus als Ausgangspunkt der erneuten Auszeichnung des Korpus.

3.1.5 Repräsentationsform

Ausgehend von der Anforderung der einfachen Benutzbarkeit werden die extrahierten Fakten im auch zur Kodierung von Ontologien verwendeten ***Resource Description Format*** (RDF) abgelegt. Damit vereinfacht sich die Handhabbarkeit der Informationen, da ein Parser für den Zugriff genutzt werden kann.

3.1.5 Repräsentationsform

Im einfachsten Fall lassen sich mit Hilfe dieser Daten beliebige Texte auszeichnen und so durch einen menschlichen Leser schneller erschließen. Interessanter ist es jedoch, die Fakten mit Hilfe der als Assoziationsregeln interpretierbaren Kommunikationsmuster in eine ontologisch modellierte Wissensbasis zu übertragen und diese so mit Inhalt zu füllen. So lässt sich ein Faktenretrieval wie bei Datenbanken realisieren. Zur Untersuchung von Ähnlichkeiten kommen Betrachtungen der Struktur sowie gemeinsamer Merkmale in Frage.

An dieser Stelle sei angemerkt, dass bei ontologischer Modellierung die Frage der Granularität stets kritisch zu sehen ist. Je höher deren Spezifität desto geringer wird die Allgemeingültigkeit. Desto umfassender der Diskursbereich sein soll, desto höher ist neben dem Aufwand der Modellierung auch der zur Abstimmung aller Beteiligten. Da ein Zuviel an Informationen eher verwirrend wirkt, sind aus Anwendersicht eher weniger Abstraktionsebenen sinnvoll.

3.2 Konzeption und Konkretisierung

3.2.1 Rahmenbedingungen

Als Domäne der prototypischen Realisierung wird der Bereich Wirtschaft und IT gewählt. Grund dafür ist das vorhandene Hintergrundwissen zur Beurteilung der Ergebnisse und zur Modellierung der Ontologie. Die zu erschließenden Informationen beruhen auf Kurznachrichten. Damit wird eine hohe Informationsdichte sichergestellt und eine einheitliche Struktur der Ausgangsdaten gewährleistet, die wenig Aufwand zur weiteren Formatierung und Aufbereitung verspricht.

Im Rahmen der Entwicklung wird der Korpus K2003, der dem Archiv des Jahres 2003 von heise online entspricht, verwendet. Fragen der Skalierung und Übertragung auf weitere Sprachen und Domänen werden an Hand der weiteren bereits in Tabelle 1 auf Seite 43 dargestellten Korpora untersucht. Zu betrachtende Sprachen sind damit Deutsch, Schwedisch und Englisch.

Die Reuterskorpora sind mit umfangreichen Metadaten des Dublin Core [DCore06] ausgezeichnet, die zur weiteren Verarbeitung entfernt wurden. Dennoch ist der englische Korpus mit vorhandener Hardware nicht zu bearbeiten. Neben der großen

Anzahl von einzelnen Dateien, die unter Microsoft Windows® nicht beherrschbar ist, stellt auch der Speicherbedarf resultierend aus den 3 GByte an Rohdaten während der Bearbeitung eine Beschränkung dar. Dieser Korpus wird daher in einen Teil Wirtschaftsnachrichten (REngC) und einen Teil sonstiger Nachrichten (REng) zerlegt.

Entwicklung und Test finden auf handelsüblicher, im Jahr 2004 aktueller Hardware mit 1 bzw. 2 GB Hauptspeicher und Prozessoren der 3-GHz-Klasse statt. Die Entwicklung erfolgte unter Microsoft Windows XP Professional. Als Programmiersprache wurde wegen der Unterstützung regulärer Ausdrücke und der effizienten Speicherverwaltung mit Hilfe von Hash-Strukturen ActivePerl® in der Version 5.8.4 verwendet.

Die Ontologie, die sowohl der Modellierung des Diskursbereichs als auch der Darstellung der Ergebnisse dient, wurde mit dem Werkzeug Protégé [Pro2000] der Stanford University in der Version 3.1.1 erstellt.

3.2.2 Beschreibung der Domäne

Wegen des prototypischen Charakters der angestrebten Umsetzung soll es genügen, einen kleinen Ausschnitt zu modellieren, der dennoch als aussagekräftig gewertet werden kann. Als Sprache zur Kodierung der Ontologie wurde wegen der weiten Verbreitung und der ausgereiften, frei zur Verfügung stehenden Werkzeuge wie Protégé das Format RDF-Schema beziehungsweise RDF [RDFS03] gewählt.

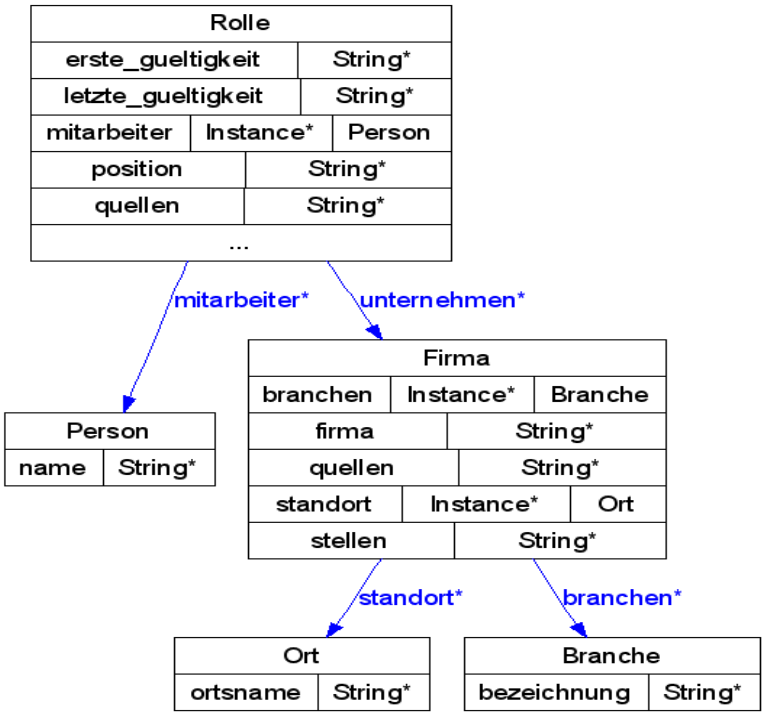
Die bekanntesten derzeit verfügbaren Ontologien mit wirtschaftlichem Hintergrund sind das ‚Ressource-Event-Agent-Model' [GeCa03] und ‚The Enterprise Ontology' [Usch+98]. Beide sind jedoch von einer hier nicht benötigten Mächtigkeit und Granularität. Eine Eigenentwicklung ist daher insbesondere zur klaren Darstellung der erreichten Ergebnisse angebracht.

Die Ontologie stellt den wichtigsten Anknüpfungspunkt zu aufbauenden Informationssystemen dar. Deren Modellierung hat, abgesehen von der Festlegung der zu extrahierenden Begriffe, vor allem konzeptionellen Charakter. Daher soll bereits an dieser Stelle und nicht erst im Rahmen der Implementierung darauf im Detail eingegangen werden.

3.2.2 Beschreibung der Domäne

Wesentliche Begriffe im Diskursbereich sind die üblicherweise betrachteten Eigennamen <FIRMA>, <PERSON> und <ORT> als konkrete Begriffe sowie die beiden eher abstrakten Begriffe <BRANCHE> und <ROLLE>. Damit lassen sich bereits wichtige Informationen wie beispielsweise die Rolle einer Person innerhalb einer Firma oder Geschäftsbereiche von Unternehmen abbilden, wobei auf Grund der zu Grunde liegenden Texte auch temporale Aspekte beachtet werden können. Abbildung 24 zeigt die Ontologie im Überblick.

Abbildung 24: Ontologie im Überblick



Quelle: eigene Erstellung

3.2.2.1 Eigennamen

Ein wesentliches gemeinsames Merkmal von Eigennamen als Bezeichnungen von Begriffen ist deren, zumindest soweit hier relevant, sprachunabhängige Großschreibung. Sie sind Gegenstand praktisch aller Systeme zur Named Entity Recognition und beides führt dazu, dass für diese Begriffe sehr gute Klassifikationsergebnisse erreicht werden. Hinsichtlich der durch die Mikromuster (Abschnitt 3.1.3.1) beschrieben Morphologie weisen diese Bezeichnungen sehr große Unterschiede auf. Daher sollen daran die Überlegungen bei der Erstellung der Mikromuster dargestellt werden.

Der Begriff ***<ORT>*** ist sowohl semantisch als auch morphologisch auf den ersten Blick am einfachsten. Er findet praktisch nur als Standort eines Unternehmens Verwendung.

Ortsnamen sind unter Umständen sehr kurz (‚Rom') und/oder bestehen aus zwei einzelnen Worten (‚Bad Ems'). In letzterem Fall stellen Verfugungen durch Leerzeichen, Binde- oder Schrägstrich die einzige Besonderheit dar. Sonstige Ortsbezeichnungen wie ‚Rothenburg ob der Tauber' oder ‚Stoke on Trend' werden auf Grund der Seltenheit nicht betrachtet. Im Fall häufiger Verwendung werden diese Formen in der Sprachpraxis ohnehin verkürzt, wie dies beispielsweise bei ‚Frankfurt/Main' der Fall ist [CaSo03].

Instanzen des Begriffs ***<PERSON>*** sind durch ein komplexeres morphologisches Muster gekennzeichnet. Neben der auf Seite 93 dargestellten Beschreibung sind vor allem im internationalen Umfeld im Wesentlichen noch weitere Abstammungsbezeichnungen zu beachten. Zum Beispiel trifft man, notiert als reguläre Ausdrücke, bei niederländischen Namen oft auf `van( de[rn]?)?` oder auf `(de la)?` im Französischen.

Diese Besonderheiten lassen sich jedoch leicht als Konjunktion in den gesamten Ausdruck einfügen. Der grundsätzliche morphologische Aufbau ist über alle hier zu betrachtenden Sprachen gleich. Eine Trennung in Vor- und Zunamen auf Ebene der Modellierung findet nicht statt, da hiervon einerseits kein Informationsgewinn zu erwarten ist und dies andererseits bei Personen anderer Kulturkreise oft sachlich nicht richtig wäre.

Die größte Komplexität in der morphologischen Beschreibung weisen die Instanzen des Begriffs ***<FIRMA>*** auf. Sie beginnen unter Umständen mit Ziffern oder Kleinbuchstaben, beinhalten Sonderzeichen und bestehen oft aus mehreren Worten. Man trifft daher auf eine Vielzahl möglicher Formen, so das kein allgemein gültiges Muster angeben werden kann. Folgende Beispiele verdeutlichen dies: ‚Siemens AG', ‚AMD', ‚Infineon Technologies', ‚1&1', ‚1&1 Internet AG', ‚123-Studios', ‚debitel', ‚mmO2', ‚et cetera', ‚Web.de', ‚Intel'. Der Grund für diese Vielfalt liegt sicher zu großen Teilen im gewählten Anwendungsgebiet, da hier Neuartigkeit und Kreativität eine wichtige Rolle spielen.

Schwerpunkt der Realisation bildet jedoch nicht die Abbildung aller morphologischen Möglichkeiten sondern die generelle Vorgehensweise. Daher wurde folgende Beschreibung angesetzt, die die unterstrichenen Fälle in obiger Aufzählung abdeckt:

> *Es können 1-2 Kleinbuchstaben oder 1 Ziffer oder 1-3 Großbuchstaben auftreten. Zwingender Bestandteil ist 1 (Großbuchstabe oder Ziffer oder Sonderzeichen) gefolgt von mindestens 1 (Ziffer oder Buchstabe). Danach folgen bis zu 2 Mal (Leerzeichen oder Bindestrich, Großbuchstabe, mindestens 1 beliebiger Buchstabe).*

Aus semantischer Sicht kommt <FIRMA> die zentrale Position innerhalb der realen Objekte der Ontologie zu. Neben der Bezeichnung sowie möglichen Alternativbezeichnungen sind weitere Merkmale die Branche und der Standort. Weiterhin gehören zu diesem Begriff über Rollen auch Personen. Bei Einbindung weiterer Aspekte wie ***<KENNZAHL>*** oder ***<VORGANG>*** würde sich die Bedeutung von <FIRMA> als zentraler Begriff weiter erhöhen.

3.2.2.2 Abstrakte Begriffe

Instanzen abstrakter Begriffe (vgl. S. 12) beginnen im Gegensatz zu den vorgenannten, typischen Eigennamen nicht sprachübergreifend mit einem Großbuchstaben. Außerdem weisen sie oft auch innerhalb einer Sprache sehr unterschiedliche Morphologien auf. Das führt dazu, dass mehrere morphologische Beschreibungen nötig sind, um alle Facetten abdecken zu können.

Beim Begriff der ***<ROLLE>*** trifft man beispielsweise auf ‚Finanzminister' als auch auf ‚Bundesminister für Finanzen'. Im Englischen findet man für die Rolle des Vorstandsvorsitzenden die Bezeichnungen ‚CEO' und ‚Chief Executive Officer', letzteres auch in Kleinschreibung.

Bezeichnungen der ***<BRANCHE>*** sind wenig spezifisch. Sie bestehen im Deutschen aus nur einem Substantiv, in anderen Sprachen oft aus ein bis zwei eventuell durch eine Präposition verbundenen Worten mit kleinem Anfangsbuchstaben. Daher sind auch hier verschiedene Typen zu definieren. In sprachlichen Gebrauch schlägt sich diese geringe Spezifität in häufiger Kompositabildung nieder. Da bei-

spielsweise ‚Software' als Branche zu unspezifisch ist, wird sie durch ‚Softwarehersteller' konkretisiert. Dieser Aspekt ist bei der Generierung der Suchmuster zu berücksichtigen.

Aus semantischer Sicht kommt der Rolle eine sehr große Bedeutung zu. Sie bildet die Verbindung zwischen <PERSON> und <FIRMA>, wobei in einer Firma mehrere Rollen existieren und eine Person mehrere Rollen in verschiedenen Unternehmen haben kann. Die Ausübung einer Rolle ist zeitlich begrenzt, so dass eine Historie mit Hilfe der Datumsinformation in den Metadaten der Texte abzubilden ist. Das erste Auftreten einer Person in einer Rolle entspricht dem Zeitstempel der ersten Meldung, die diese Information enthält. Tritt eine andere Person in ebendieser Rolle auf, so ist der vorhergehende Tag der letzte gültige Zeitpunkt.

3.2.3 Die Extraktionskomponente

3.2.3.1 Morphologische Beschreibungen

Wie dargestellt soll die Erkennung neuer Begriffsinstanzen mit Hilfe der Makromuster geschehen, wobei jeweils ein Begriff durch seine morphologische Beschreibung, sein Mikromuster, zu ersetzen ist. Das bedeutet, dass vorab zu definieren ist, welcher Morphologie ein Begriff gehorchen soll.

Da die extrahierten Fakten in einem XML-ähnlichen Format abgelegt werden sollen, liegt es nahe, die Morphologie als ein Attribut des jeweiligen Begriffs zu kodieren. Während für die herkömmlichen Eigennamen jeweils eine Beschreibung ausreichend ist, sind für abstrakte Begriffe oft unterschiedliche Beschreibungen notwendig. Entsprechend ist für jeden zu identifizierenden Begriff einer morphologischen Ausprägung mindestens ein Beispiel im Seed (siehe S.95) anzugeben.

Schwieriger gestaltet sich unter Umständen die Definition des Mikromusters an sich. Neben der hier verfolgten manuellen Vorgehensweise wie sie oben dargestellt wurde, wäre auch ein Lernen aus Beispielen möglich (vgl. [Röss04, CoSi99]). Allerdings setzt diese Vorgehensweise ebenfalls viel Wissen über die auftretenden Wortformen voraus, um aussagekräftige Beispiele angeben zu können. Würden als Beispiele für Firmenbezeichnungen nur 'Siemens', 'Infineon' und 'Bosch' genannt,

so bestünde keine Möglichkeit, daraus auf ‚1&1 Internet AG' als gültiger Bezeichnung zu schließen.

Ein weiterer Aspekt der Mikromuster ist hinsichtlich unterschiedlicher Zeichensätze relevant, beispielsweise bei Übertragung auf Korpora anderer Sprachen. Da die hardwareseitige Unterstützung dieser Spezifika nicht gewährleistet ist, müssen auch Sonderzeichen Bestandteil der Mikromuster sein. Ein Ansatz wäre dazu die explizite Vereinbarung, welche Zeichen als Klein- bzw. Großbuchstaben zulässig sind.

3.2.3.2 Mustergewinnung

Mit Hilfe der anfänglich gegebenen Instanzen wird der Korpus ausgezeichnet, indem diese durch die jeweiligen, in der Ontologie definierten Begriffe ersetzt werden. Diese werden durch die dort vereinbarte, in spitze Klammern gefasste Hauptbezeichnung repräsentiert (z.B. ***<PERSON>***).

Anschließend werden alle Textausschnitte selektiert, die von einer solchen Ersetzung betroffen sind. Zur Identifikation der Kommunikationsmuster ist es notwendig, dass die Textausschnitte eine gewisse Länge aufweisen und der betrachtete Begriff eher zentral darin auftaucht. Daher wird ausgehend von der Position eines Begriffs ein Bereich vor- und nachher selektiert. Die Größe dieses Bereichs muss einerseits auch Muster mit mehr als zwei Begriffen fassen können und andererseits Muster vermeiden, bei denen der Abstand zwischen den Begriffen zu groß und deren Bindung im Sinne einer Kollokation damit zu gering ist. Eine Größe von je 40 Zeichen hat sich dabei als günstig insbesondere auch zur Identifikation von Mustern aus drei und mehr Begriffen erwiesen.

Je nach Anzahl der enthaltenen Begriffe werden die Textausschnitte auf Grundmuster reduziert. Da wie auf Seite 95 dargestellt die Aussagekraft eines Musters mit der Zahl der beteiligten Begriffe steigt, sind unterschiedliche TrustLevel (TL) zu unterscheiden. Muster des TL1 entsprechen einem durch Freitext eingeschlossenen Begriff und alle höheren Muster repräsentieren Begriffe mit dadurch eingeschlossenem Freitext. Deren prinzipiellen Aufbau zeigt Abbildung 25. Im einzelnen werden nur Muster mit bis zu drei Begriffen betrachtet, deren Verarbeitung jedoch auch für alle höheren TrustLevel typisch ist.

Abbildung 25: Prinzipieller Aufbau der Makromuster nach TrustLevel

- **TL3:** <BEGRIFF>...<BEGRIFF>...<BEGRIFF>
- **TL2:** <BEGRIFF>...<BEGRIFF>
- **TL1:** ...<BEGRIFF>...

Anschließend erfolgt eine erste Bereinigung der Muster, indem beispielsweise Grundmuster mit nur einem Begriff auf einen eventuell enthaltenen, durch Komma-ta identifizierbaren Teilsatz reduziert werden. Außerdem werden zu seltene Muster gelöscht, da sonst leicht eine Unregelmäßigkeit als gültiges Kommunikationsmuster betrachtet wird.

Weiterhin sind Muster zu vermeiden, die zu Fehlklassifikationen führen. Eine wesentliche Ursache ist dabei der Gebrauch der Worte in der natürlichen Sprache, der zu unterschiedlichen Interpretationsmöglichkeiten führt. Insbesondere Bezeichner abstrakter Begriffe bestehen häufig aus Generalisierungs-Spezialisierungs-Relationen (GSR). Beispielsweise spezialisiert ‚Sprecher' die generelle Rolle ‚Vorstand' in der Bezeichnung ‚Vorstandssprecher'. Generalisierende Wortbestandteile sind gute Kandidaten für eingenständige Begriffe bzw. deren typische Bezeichner. Im Beispiel wäre ‚Vorstand' der Oberbegriff für ‚Vorstandssprecher', ‚Vorstandsvorsitzende', etc.

Tritt in einem potenziellen Muster der generalisierende Teil als Freitext auf, so führt dies leicht zu Fehlern. Der Ausschnitt ‚...Microsoft-Sprecher...', resultierend in dem Makromuster ‚Microsoft-<ROLLE>', würde beispielsweise ‚Aktie' und ‚Produkt' ebenfalls als Instanzen von <ROLLE> klassifizieren, falls im Text ‚Microsoft-Aktie' oder ‚Microsoft-Produkt' vorkommen.

Die Frage, was generell und was spezifisch ist, hängt im Gegensatz zum üblichen Sprachgebrauch, vom Anwendungsfall ab. Interessierten die verschiedenen Arten von Sprechern wäre die Relation entgegengesetzt. Da die zu Grunde liegende Dokumentenmenge unter Anwendungsaspekten gewählt wird, könnte als Anhaltspunkt dafür, was generalisierend und was spezialisierend ist, die Häufigkeit herangezogen werden, mit der ein Wort einzeln und im Zusammenhang auftritt. Im Korpus K2003 beispielsweise tritt das Wort ‚Microsoft' 4420 mal auf und Sprecher 845 mal. Bei ‚Microsoft-Sprecher' wäre entsprechend ‚Sprecher' der spezialisierende Teil.

Muster, deren Freitextteil einen bisher nicht klassifizierten Eigennamen aufweisen sind also nicht zur Erschließung von Begriffsbezeichnern geeignet und werden gelöscht. In Ermangelung anderer Kriterien kann zur Beurteilung dazu, was ein potentieller Eigenname ist, nur eine statistische Vorgehensweise in Betracht gezogen werden.

3.2.3.3 Statistische Analyse

Zur Auswahl geeigneter Kenngrößen werden die Untersuchungen aus Kapitel 2.2.5 herangezogen. Über alle Korpora hinweg zeigte sich, dass die Häufigkeit eines Wortes wenig aussagekräftig ist und Worte mit einer Auftretenswahrscheinlichkeit größer 1/300 im Allgemeinen gute Kandidaten für Stoppworte darstellen. Unterwürfe man den Korpus vorab einer syntaktischen Analyse mit Grundformenreduktion, wäre das Ergebnis sicher ein anderes, da sich auf Grund der Reduktion auch die Häufigkeiten ändern.

Ziel der Relevanzschätzung ist es, häufige und gleichzeitig zu spezifische Muster zu vermeiden, da diese wie dargestellt leicht zu Fehlklassifikationen führen können. Daher wird sowohl für einzelne Worte wie auch für 2-Grams die Kenngröße TF-IDF verwendet. Letztere werden nur beachtet, wenn sie kein hochfrequentes Wort ($p > 1/300$) beinhalten, und beide Teile jeweils mit einem Klein- oder einem Großbuchstaben beginnen. Dies erwies sich über alle Korpora als empirisch nachvollziehbare, sinnvolle Einschränkung. Die Identifikation längerer Kombinationen stellte sich im Rahmen vorgelagerter Untersuchungen angesichts des erforderlichen Aufwands als wenig sinnvoll heraus, da für jede auftretende Kombination von Worten zu untersuchen ist, wie häufig diese ist. Selbst bei kleinen Korpora stößt man somit schnell an die Grenzen des derzeit technisch Realisierbaren, dehnt die Laufzeit stark aus oder benötigt effizientere Algorithmen zur Speicherverwaltung, die jedoch nicht Gegenstand dieser Arbeit sind.

Großschreibungen von Satzanfängen führen leicht zu Mustern, die prinzipiell auch innerhalb eines Satzes auftreten, aber wegen der Schreibweise nicht gefunden werden. Daher wird untersucht, welche Worte sowohl in Klein- wie auch Großschreibung vorkommen und welche Form die häufigere ist. Mit Hilfe dieser Information lassen sich identifizierte Muster gegebenenfalls generalisieren. Im Rahmen des

folgenden Abschnitts 3.2.3.4 wird die Anwendung dieser Informationen an einem Beispiel verdeutlicht.

Um nicht bei jedem Lauf die gesamten statistischen Daten erneut ermitteln zu müssen, werden die Ergebnisse des ersten Laufs gesichert und später nur hinzu geladen. Eine Prüfung, ob der zu bearbeitende Korpus zu den evtl. vorhandenen Statistikdaten korrespondiert, erfolgt nicht. Dies ließe sich aber leicht mit Hilfe eines Prüfsummenverfahrens realisieren.

3.2.3.4 Aufbereitung und Filterung der Muster

Die Anwendung der identifizierten Kommunikationsmuster zur Extraktion neuer Begriffsinstanzen verursacht den höchsten Aufwand der gesamten Lernphase, da jedes Muster im gesamten Korpus zu suchen ist. Daher kommt der Selektion erfolgversprechender Muster große Bedeutung zu. Entsprechend des Ansatzes, die Makromuster als feste Folgen von Symbolen und Variablen abzubilden, spielen neben den oben genannten Kriterien in erster Linie strukturelle Merkmale der potentiellen Muster eine Rolle.

Muster mit nur einem beinhalteten Begriff (TL1) sind am kritischsten. Sie treten vor allem zu Beginn der Lernphase bei noch wenigen bekannten Begriffsinstanzen sehr häufig auf. Als strukturelle Merkmale kommen die Länge, die Wortzahl und die Verteilung bezüglich des Begriffs in Frage. Befindet sich nur auf einer Seite des Begriffs Freitext, werden Länge und Wortzahl noch wichtiger.

Bei Mustern des TL1 werden vom Begriff ausgehend in beide Seiten solange Worte akzeptiert, wie deren Länge zumindest nicht abnimmt. Hintergrund ist hier wiederum, dass längere Worte seltener und somit spezifischer sind. Aus dem Ausgangsmuster ‚`Sehr deutlich sagte <PERSON> gestern bei einer`‘ wird dementsprechend ‚`deutlich sagte <PERSON> gestern`‘.

Wird bei der Anwendung der Ergebnisse aus der statistischen Analyse festgestellt, dass das erste Wort sowohl in Klein- wie auch Großschreibung möglich ist, wird das Muster zu ‚`[dD]eutlich sagte <PERSON> gestern`‘ generalisiert. Nur Muster, die während dieser Reduktion nicht zu stark gekürzt werden, kommen für die weitere Verarbeitung in Betracht.

Seien als Beispiel die beiden Muster ‚neue <FIRMA>-Chef' sowie ‚ebenfalls <FIRMA>' betrachtet. Beide bestehen inklusive Leerzeichen aus 10 Zeichen. Da im ersten Fall der Begriff jedoch eingebettet ist und eine Verbindung ungleich Leerzeichen zu weiteren Zeichen hat, würde es weiterhin als gültig betrachtet, das zweite Muster hingegen gelöscht.

Bei Mustern des TL2 spielen neben der Länge des Freitextes die betrachteten Begriffe eine wichtige Rolle. Ein Indiz für ungültige bzw. ungeeignete Muster liegt vor, falls beide Begriffsbezeichner den gleichen Begriff repräsentieren. Diese werden gelöscht.

Vor dem Hintergrund des verwendeten Begriffsverständnisses würde ein Muster, welches zweimal den gleichen Begriff beinhaltet bedeuten, dass ein Begriff durch Bezugnahme auf sich selbst definiert wird. Dies ist nur im Fall von Aufzählungen möglich, wobei eine zweistellige Aufzählung praktisch nicht vorkommt. Eine weitere Erklärung solcher Fälle sind neben Fehlklassifikationen vor allem unterschiedliche Abstraktions- oder Generalisierungsebenen, die auf einen einzigen Begriff Bezug nehmen. Ein Beispiel dafür ist der Textausschnitt ‚Ein Sprecher des neuen Vorstands teilte'. Die Bezeichner Sprecher und Vorstand bezeichnen beide den Begriff <ROLLE>, allerdings auf verschiedenen aber nicht unterschiedenen Generalisierungsebenen.

Die geringsten Anforderungen stellen Muster aus drei oder mehr Begriffen. Diese geben im wesentlichen Aufzählungen oder Zugehörigkeiten von Begriffen zueinander wieder, die ihrerseits zur Verifikation oder Erweiterung der Ontologie genutzt werden können. Nur falls ein Begriff genau zweimal auftritt, ist dies wie bei Mustern des TL2 ein Indiz für einen Fehler. In diesem Fall werden auch Muster des TrustLevel 3 gelöscht.

3.2.3.5 Bewertung und Generalisierung

In die abschließende Bewertung der Muster fließen die morphologische Komplexität der beteiligten Begriffe sowie die Anzahl und Länge der Worte im Freitext ein. Insbesondere bei Mustern des TrustLevel 2 ist zu beachten, dass morphologisch unspezifische Beschreibungen wie sie bei abstrakten Begriffen vorkommen, die

Verlässlichkeit des Musters verringern. Wird beispielsweise aus der Textstelle ‚`wie ein Sprecher des französischen IT-Dienstleisters mitteilte`' das Muster ‚`<ROLLE> des französischen <BRANCHE>`'. Dies führt dazu, dass an der Textstelle ‚`die Übernahme des französischen Konkurrenten`' ‚Übernahme' als <ROLLE> und ‚Konkurrenten' als <BRANCHE> klassifiziert werden. Wie sich im Rahmen von Voruntersuchungen herausstellte, tritt eine solche Fehlklassifikation bei morphologisch komplexen Begriffen wie beispielsweise <PERSON> oder <FIRMA> nur sehr selten auf. Entsprechend ist diesen Mustern bei morphologisch einfachen Begriffsbezeichnern ein geringeres Gewicht zuzuordnen.

Diese Gewichtung wird ergänzt durch die Häufigkeit des jeweiligen Musters sowie dessen TrustLevel. Um bisher selten identifizierte Instanzen eines Begriffs gezielt erschließen zu können, wird ein Dringlichkeitsfaktor *force* gemäß Formel 12 ermittelt, der mit zunehmender Anzahl identifizierter Instanzen eines Begriffs, bezeichnet als *anz(con$_i$),* abnimmt.

Formel 12: Berechnung des Dringlichkeitsfaktors eines Begriffs

$$force(con_i) = e^{-0.01*anz(con_i)}$$

Die Signifikanz *Sig* eines Musters berechnet sich gemäß Formel 13. Dabei bezeichnet *øforce* die durchschnittliche Dringlichkeit der beinhalteten Begriffe, *tl* den TrustLevel des Musters, *anz* dessen Häufigkeit und *gewicht* die gewichtete Länge des enthaltenen Freitextes.

Formel 13: Signifikanz des i-ten Musters

$$Sig = \varnothing\text{force} * \log(2*\text{tl}*(\text{tl}+3) + \text{anz} + \text{gewicht})$$

Um *tl* einen höheren Stellenwert beizumessen, werden die auftretenden Werte 1, 2, 3... zu 8, 20, 36... transformiert. Die Summe der einfließenden Werte wird logarithmiert, um extreme Werte zu relativieren. Über alle in der Evaluation dargestellten sowie weitere Programmläufe zeigte sich Formel 13 als geeignet. Die Skalierung des TrustLevel ist variierbar, sollte aber im dargestellten Bereich liegen.

Die bewerteten Muster werden gemäß der erreichten Signifikanzwerte sortiert. Als Grenzwert zur weiteren Verarbeitung wird der Wert des 2/3-Perzentils genutzt. In die im Folgenden beschriebene Generalisierung fließen Muster ein, deren Bewertung besser als das 3/4-Perzentil ist. Insgesamt werden jedoch maximal 50 bzw. 55 Muster zur Anwendung zugelassen.

Zur Erhöhung des Recalls werden bis hierhin akzeptierte Muster aus zwei Begriffen aufgeteilt, falls es eine direkte Verbindung zwischen einem Begriff und dem daran angrenzenden Freitext gibt und das resultierende Muster wie im vorigen Abschnitt dargestellt, hinreichend lang ist. Aus dem Muster ‚`<FIRMA>-Vorstandschef <PERSON>`' wird so zusätzlich das Muster ‚`<FIRMA>-Vorstandschef`' abgeleitet.

Die abschließende Generalisierung dient der Erhöhung des Recall bei gleichzeitiger Beschleunigung der Musteranwendung, wie das folgende Beispiel zeigt:

Die Muster `<BRANCHE>-Hersteller` und `<BRANCHE>herstellers` lassen sich zu `<BRANCHE>(-H|h)ersteller[s]?` generalisieren und decken mit nur einem einzigen Anwendungszyklus neben den beiden ursprünglichen beispielsweise auch das Muster `<BRANCHE>hersteller` ab. Realisiert wird die Generalisierung durch Bestimmung der längsten gemeinsamen Zeichenfolge und ist in Abschnitt 3.3.3 genauer dargestellt.

3.2.3.6 Anwendung und Prüfung

Die ermittelten Muster werden um Muster aus früheren Zyklen, die in der Regelbasis vorhanden sind, ergänzt. Treten dabei nicht zu lösende Widersprüche auf, so ist dem älteren Muster Vorrang zu geben. Grund dafür ist die Erwartung, dass im Laufe des Lernprozesses die Zahl von Fehlern zunimmt und so ungültige Muster wahrscheinlicher werden. Unter einem Widerspruch ist dabei zu verstehen, dass im Prinzip identische Muster unterschiedliche Begriffe beinhalten.

Mit den Extraktionsmustern wird der Korpus durchsucht, indem für jeden enthaltenen Begriff das entsprechende Mikromuster eingesetzt wird. Trifft das Muster auf eine Textstelle zu, lässt sich an der durch das Mikromuster beschriebenen Position ein neuer Bezeichner dieses Begriffs klassifizieren. Ergänzend zur Klassifikation der Instanzen wird dabei angewandte Muster ebenso protokolliert wie weitere

Klassifikationen desselben Bezeichners. Diese Informationen dienen der Validierung der Muster und der durchgeführten Klassifikation.

Im ersten Schritt der Prüfung wird untersucht, ob eine Instanz so klassifiziert wurde, dass sie ein Makromuster entsprechend der Filterkriterien (vgl. Abschnitt 3.2.3.4, S.105) ungültig machen würde. In diesem Fall sind die beteiligten Muster sowie alle dadurch klassifizierten Bezeichner zu löschen bzw. als nicht klassifiziert zu betrachten. Alternativ könnten diese Bezeichner als latent klassifiziert betrachtet und nach weiteren Hinweisen zur Entscheidung gesucht werden.

Der zweite Prüfungsschritt widmet sich durch Ambiguitäten bzw. Verallgemeinerungen der natürlichen Sprache hervorgerufenen Mehrfachklassifikationen. Diese werden in der natürlichen Sprache oft benutzt, um Wiederholungen von Worten zu vermeiden und bilden einen Ansatz zu der im folgenden Abschnitt dargestellten ontologischen Betrachtung. Um nicht wiederholt sagen zu müssen ‚wie Klaus Petersen mitteilte', wird beispielsweise ‚wie Siemens mitteilte' benutzt, obwohl ‚Siemens' eigentlich eine <FIRMA> ist. Die dabei verfolgte Vorgehensweise zur Auflösung dieser Konflikte ist in Abschnitt 3.3.5.2 im Detail dargestellt.

Die am Ende als gültig betrachteten Makromuster werden der Regelbasis und die Bezeichner der Faktenbasis hinzugefügt. Beides zusammen bildet die Schnittstelle zu einem darauf aufbauenden Informationssystem, dient als Ausgangspunkt weiterer Lernzyklen und ergänzt die Schnittstelle zur Ontologie.

3.2.4 Ontologische Betrachtung

Der Gebrauch der natürlichen Sprache ist auch bei Fachtexten durch Ambiguitäten, Unregelmäßigkeiten und auf Hintergrundwissen aufbauende Auslassungen gekennzeichnet. Beispielsweise werden aus stilistischen Gründen oft Gattungs- an Stelle von Eigennamen genutzt, um Wiederholungen zu vermeiden. Um dieser Problematik zu begegnen, sind Oberbegriffe zu identifizieren.

Die Grundidee dabei ist entsprechend der oben dargestellten Generalisierungs-Spezialisierungs-Relationen (siehe S.103), dass Gattungsnamen sehr oft zur genaueren Spezifikation von Eigennamen genutzt werden. Tritt also ein potentieller

Eigenname in vielen Verbindungen mit anderen Eigennamen auf, so handelt es sich wahrscheinlich um einen Gattungsbezeichner.

Dabei sind zwei Fälle zu unterscheiden. Tritt eine solche Gattungsbezeichnung innerhalb eines Begriffs auf, so handelt es sich um ein Hyperonym. Ein Beispiel innerhalb des Begriffs <ROLLE> ist `Vorstand`, wenn gleichzeitig `Vorstandssprecher`, `Vorstandsvorsitzender` und `Finanzvorstand` auftreten.

Im anderen Fall tritt eine Bezeichnung als Bestandteil von Bezeichnern eines anderen Begriffs auf. Ein typisches Beispiel hierfür ist `Konzern` als <FIRMA> und als Bestandteil vieler Instanzen bspw. des Begriffs <BRANCHE>, wie in `Halbeiterkonzern` oder `Microsoft-Konzern`. Dies kann so interpretiert werden, dass zum einen ‚Konzern' eine allgemeine Bezeichnung im Begriff <FIRMA> darstellt, also entsprechend der hier verwendeten Systematik (siehe S.12) selbst einen Begriff bezeichnet. Andererseits weist die 1:n-Beziehung darauf hin, dass <BRANCHE> attributiv zu <FIRMA> gehört.

Dieses Beispiel zeigt einen weiteren Ansatz zur Beseitigung von Fehlklassifikationen auf. Tritt die erwähnte klassenübergreifende Verbindung selten auf, wie es bei `Microsoft` als <FIRMA> und `Microsoftkonzern` bzw. `Microsoft-Konzern` als <BRANCHE> der Fall ist, so kann man davon ausgehen, dass letztere falsch klassifiziert wurden. Die Umsetzung dieser Überlegungen zur Ontologie zeigt Abschnitt 3.3.6.

Bevor die Konzeption durch den Gesamtüberblick und die Evaluationsplanung ihren Abschluss findet, sollen noch einige, für die Konzeption relevante Aspekte beleuchtet werden. Diese sind zwar nicht von zentraler Bedeutung für die Vorgehensweise an sich, zeigen aber interessante Erweiterungsmöglichkeiten auf.

3.2.5 Ergänzende Aspekte

3.2.5.1 Bindestriche

Sehr fehleranfällig ist der Gebrauch des Bindestrichs, der inkonsistent zur Verbindung oder Trennung von Worten genutzt wird. Insbesondere bei abstrakten Begriffen in Generalisierungs-Spezialisierungs-Relationen ist dies der Fall. Wird zum

Beispiel das Muster `<ROLLE>-Sprecher`, nicht jedoch `<ROLLE>sprecher` identifiziert, fände keine Generalisierung statt und entsprechend weniger Instanzen würden identifiziert werden.

Daher wird bei Begriffen einfacher Morphologie beides als richtig unterstellt und die Schreibweise mit Bindestrich stets mit der Schreibweise ohne Bindestrich verschmolzen. Bei solchen Mustern ist zudem der gesamte Ausdruck als Begriffsinstanz zu betrachten, da der eigentliche Begriffsteil allein zu unspezifisch ist, wie das Beispiel zur Strukturgröße auf Seite 94 nahe legt. Die Behandlung von Bindestrichen wird in der prototypischen Umsetzung integriert.

3.2.5.2 Feste Symbole

Bei einigen Begriffen wie beispielsweise Wochentagen sind nur sehr geringe Ausbeuten durch Lernen zu erwarten, da ihre Instanzen in einer Sprache einen festen Wertebereich haben. Daher ist es probat, diese als feste Symbole in der Ontologie abzulegen. Ein Muster wie `gestrigen <wochentag> in <ORT>` tritt eher hervor und erreicht höhere Recallwerte als `gestrigen Montag in <ORT>`[22].

Eine ähnliche Vorgehensweise ist auch für Beträge, Ordnungs- und Jahreszahlen denkbar, die einer typischen Morphologie folgen. Ersetzt man diese Instanzen durch eine Begriffsbezeichnung, wären ebenfalls Muster eher erkennbar.

Allerdings geht damit eine Einschränkung der Sprachunabhängigkeit einher, so dass für jede in Frage kommende Sprache die entsprechenden Vorgaben zu treffen sind. In einem tatsächlichen Anwendungssystem spielt diese Einschränkung eine untergeordnete Rolle. Wegen des zu erwartenden Nutzens ist der geringe zu erbringende Aufwand gerechtfertigt.

3.2.5.3 Synonyme

Viele Bezeichner wie insbesondere Namen von Personen zeigen eine große Varianz. So bezeichnen 'Craig Barrett', 'Craig R. Barrett' und 'Craig R. Baret' eine einzige Person. Ein ähnliches Problem tritt bei Firmen auf, wobei hier in Abhängigkeit vom Kontext Auslassungen stattfinden. In einem deutschen Text trifft man bei-

22 Die Kleinschreibung von <wochentag> bedeutet, dass dieser Begriff kein Klassifikationsziel ist.

spielsweise nur sehr selten auf 'Deutsche Telekom AG' aber sehr oft auf 'Telekom'. Insbesondere die Auslassung der Rechtsform ist sehr üblich.

Diese Synonymbeziehungen sowie kleinere Rechtschreibfehler lassen sich in vielen Fällen mit Hilfe eines phonetischen Hashings erkennen. Hierzu kommt der in vielen Programmiersprachen standardmäßig enthaltene Soundex-Algorithmus in Frage. Dieser realisiert eine Abbildung eines Wortes auf einen meist dreistelligen Zahlenwert entsprechend des Klanges. Ausgelassene Worte wie im Beispiel 'Deutsche' sind damit natürlich nicht zu erfassen.

3.2.5.4 Abkürzungen

Eine besondere Form von Synonymen sind Abkürzungen. Diese werden oftmals eingeführt, indem die verwendete Abkürzung in Klammern eingeschlossen der Langbezeichnung folgt. Allerdings entsprechen die gegebenen Erklärungen nicht unbedingt der Buchstabenfolge in den Abkürzungen wie das Beispiel '...Ergebnis vor Zinsen, Steuern, und Abschreibungen (EBITDA)' zeigt. Eine 1:1-Zuordnung von Buchstaben zu Worten ist daher nicht sinnvoll.

Sehr gute Ergebnisse liefert eine Vorgehensweise, die von einer in Klammern stehenden Abkürzung ausgehend die davorstehenden Worte positionsweise den Buchstaben der Abkürzung zuordnet. Damit lässt sich neben dem erwähnten Beispiel auch ‚VG Wort' als ‚Verwertungsgesellschaft Wort' oder ‚Ust-IdNr.' für ‚Umsatzsteuer-Identifikationsnummer' erkennen.

3.2.5.5 Zitate

Insbesondere bei dem hier vorrangig betrachtetem Texttyp der Nachrichten spielen Zitate eine wichtige Rolle. Sie befinden sich an Stellen, an denen sehr treffend und knapp Zusammenhänge und Meinungen dargestellt werden und geben oftmals schwer zu formalisierende Auslöser für Veränderungen wieder. Deswegen sollten Zitate je nach Anwendungsfall gesondert betrachtet werden.

3.2.5.6 Integration

Zur Unterstützung der Offenheit und Erweiterbarkeit der Ontologie bzw. des darauf aufbauenden Informationssystems wäre es denkbar, die englischen Bezeichnungen der Klassen mitzuführen. Mit Hilfe der Ontologie WordNet [Mill+03] lassen sich so Verbindungen zu englischsprachigen Systemen erreichen.

Die begriffliche Erschließung bietet weiterhin einen Ansatz zur Identifikation von in einem Text behandelten Themen (***Topic Identification***). Ähnlich der wortbasierten Vorgehensweise (siehe S.34) spricht das Auftreten bestimmter Begriffe für ein spezifisches Thema. Je nach Szenario lässt sich damit eine passende Ontologie zur Strukturierung der Informationen wählen.

3.2.6 Gesamtüberblick

In Abbildung 26 ist das Zusammenspiel der Systemkomponenten im Überblick dargestellt. Anfänglich sind die morphologischen Beschreibungen der interessanten Begriffe (1) und einige wenige Beispiele interessierender Begriffe gegeben, die als ***Seed*** die Faktenbasis (2) darstellen.

Mit Hilfe eines Parsers wird die Faktenbasis eingelesen und der Korpus mit den korrespondierenden Begriffen ausgezeichnet (a). Anschließend werden diejenigen Textausschnitte selektiert, in denen eine solche Ersetzung auftrat. Danach erfolgt wie in Abschnitt 3.2.3.2 dargestellt die Reduktion der Ausschnitte auf die Begriffe und deren Umfeld gemäß TrustLevel (b).

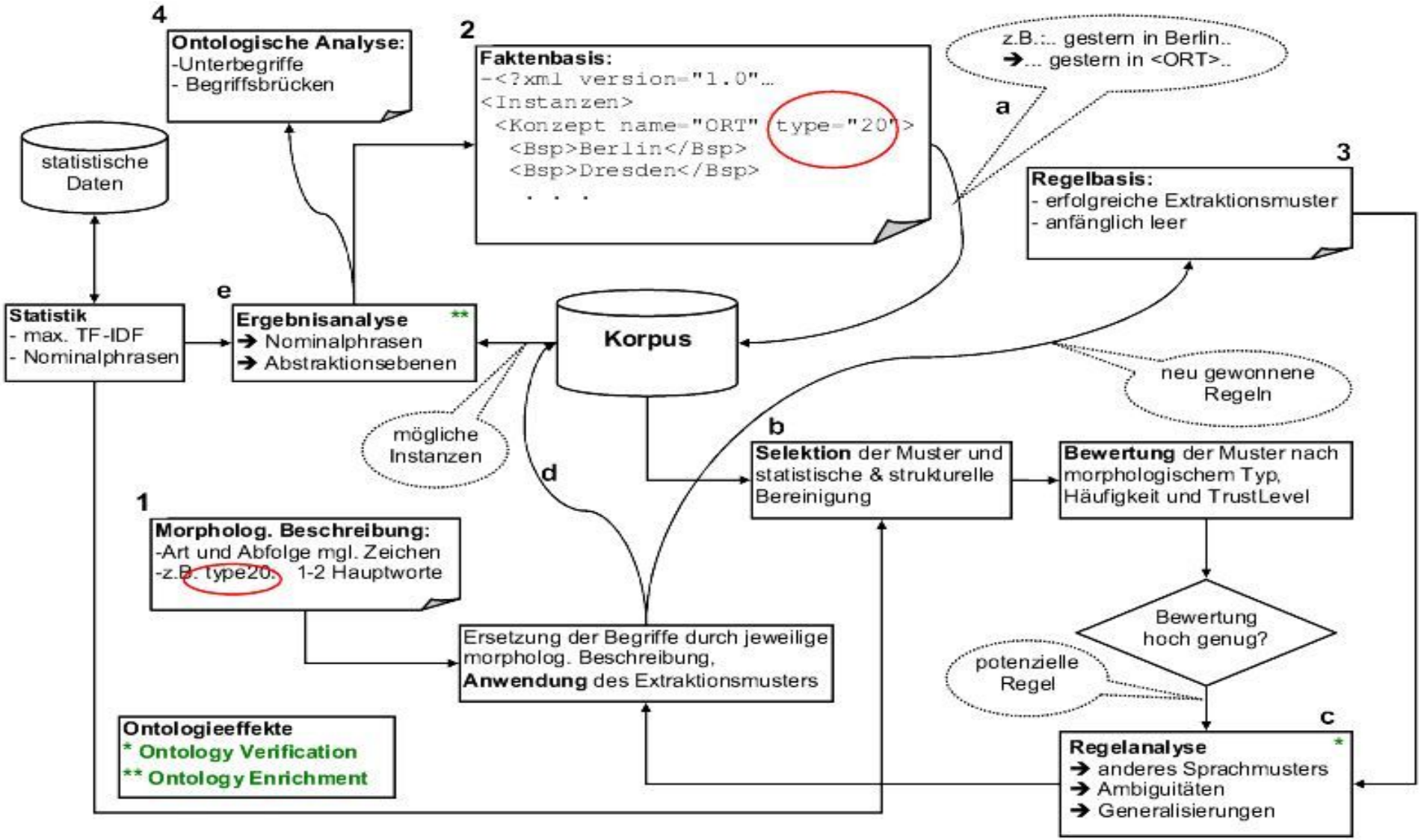
Nach der Selektion und Bereinigung der Muster erfolgt deren Bewertung und Generalisierung (vgl. Abschnitt 3.2.3.5) und anschließend werden, soweit vorhanden, Muster aus früheren Läufen aus der Regelbasis (3) dazu geladen und hinsichtlich Konflikten untersucht (c). Hier bietet sich als Erweiterung ein Ansatz zum ***Ontology Verification***.

Zur Anwendung der gewonnenen Muster werden schrittweise alle enthaltenen Begriffe durch deren jeweiliges Mikromuster ersetzt und der Korpus damit durchsucht (d). Die dabei identifizierten Bezeichner werden bezüglich ihrer Klassifikation in der Ergebnisanalyse (e) überprüft und ergänzen bei erfolgreicher Prüfung die Faktenbasis. Die erfolgreich angewandten Regeln erweitern die Regelbasis und stehen für

weitere Programmläufe sowie darauf aufbauende Untersuchungen zur Verfügung. Außerdem fließen sie in die ontologische Analyse (4), die technisch Bestandteil von (e) ist und eine Möglichkeit zum ***Ontology Enrichment*** darstellt.

Um das Lerntempo steuern und damit Fehlklassifikationen vermeiden zu können, erscheint es sinnvoll, zwischen vollständigen Lernvorgängen und Vorgängen, die lediglich bekannte Muster anwenden, zu unterscheiden. Dazu kann dem Programm ein entsprechender Parameter mitgegeben werden. Entsprechend wird später in ersterem Fall auch von einem Lernschritt bzw. ***Learn*** und im zweiten Fall von einem Anwendungsschritt bzw. ***Reuse*** gesprochen. Die aufeinander folgende Durchführung von ***Learn*** und ***Reuse*** kann zur Bestimmung einer geeigneten ***Lernstrategie*** beispielsweise in Abhängigkeit vom zu Grunde liegenden Korpus genutzt werden. Dieser Aspekt wird im Rahmen der Evaluation wieder aufgegriffen.

Abbildung 26: Zusammenspiel der einzelnen Komponenten



3.2.7 Evaluationsplanung

Wesentlichstes Ziel der Evaluation soll eine Aussage darüber sein, inwieweit der vorgestellte Ansatz in der Lage ist, die gesteckten Erwartungen unter den gegebenen Voraussetzungen zu erfüllen. Diese sind:

- Sprachneutralität durch Verzicht auf Part-of-Speech-Tagging sowie grammatische Regeln. Dabei bedeutet Sprachneutralität nicht, dass das Programm ohne Anpassung für andere Sprachen einsetzbar ist. Ziel ist vielmehr eine Vorgehensweise, die eine sprachabhängige Vorverarbeitung obsolet macht. Auf POS-Tagging aufbauende Verfahren sind dadurch gekennzeichnet, dass der eigentlichen Klassifikationskomponente ein sprachlich neutralisierter Input geliefert wird, wobei zumindest teilweise bereits eine Klassifikation von Worten als Eigenname stattfindet.
- Geringer manueller Aufwand durch Lernen aus einer sehr geringen Anzahl von Beispielen ohne vorab modellierte Extraktionsregeln. Damit genügt eine allgemeine Kenntnis der im Anwendungsbereich behandelten Entitäten, so dass die Erstellung geeigneter Beispielmengen vereinfacht wird. Ein manuell aufbereiteter Trainingskorpus ist nicht notwendig.

Im Einzelnen sollen folgende Fragestellungen genauer untersucht werden:

- Welche Werte bezüglich Precision, Recall und f-Measure erreicht der Ansatz am Entwicklungskorpus K2003, wie sieht dazu die optimale Lernstrategie, also die Abfolge von Lern- und Anwendungsschritten, aus und welcher zeitliche Aufwand ist zu erbringen?
- Wie skaliert das Verfahren bei Anwendung am Korpus K9704? Neben zeitlichen Aspekten interessiert hier auch die Frage, ob umfangreichere Korpora zu besseren Ergebnissen führen und ob sich die optimale Lernstrategie ändert.
- Was lässt sich über die Auswirkungen falsch klassifizierter Instanzen sagen? Lassen sich Verallgemeinerungen auf begrifflicher Ebene, beispielsweise zu nicht vorab definierten Begriffen, treffen?
- Welche Auffälligkeiten und unerwarteten Ergebnisse treten auf? Lassen sich Aussagen zu sinnvollen Erweiterungen ableiten?

3.2.7 Evaluationsplanung

- Der NIST/Reuters Korpus in Deutsch (RDeu) beinhaltet Nachrichten zu unterschiedlichsten Themen mit einem Schwerpunkt auf Politik und Wirtschaft. Daher soll an diesem untersucht werden, wie sich eine Ausdehnung des zuvor eng eingegrenzten Diskursbereichs auswirkt.
- Die grundsätzliche Übertragbarkeit des Ansatzes auf andere Sprachen ist am schwedischen (RSve) und englischen (REngC) Korpus der NIST/Reuters-Korpora ([Reut05] bzw. [Reut00]) zu untersuchen. Dabei interessieren neben den Kenngrößen vor allem notwendige Anpassungen am Programm

Als Vergleichsgröße bezüglich Precision und Recall werden die veröffentlichten Ergebnisse der ***Conference on Computational Natural Language Learning*** [CoNLL03] heran gezogen. Diese Ergebnisse beziehen sich auf den englischen Reuters-Korpus sowie einen deutschsprachigen Korpus der Frankfurter Rundschau, der hier nicht zur Verfügung steht. Auf Grund der unterschiedlichen Herangehensweise sind direkte Vergleiche nur für die konkreten Begriffe <PERSON>, <ORT> und <FIRMA> möglich. Damit erlauben die bei [CoNLL03] angegebenen Ergebnisse nur einen Vergleich der Größenordnung, was für das angestrebte Ziel der Evaluation ausreichend ist. Vergleichsdaten für Schwedisch sind nicht bekannt.

Zur Ermittlung der Retrieval-Kenngrößen Recall, Precision und f-Measure der beiden Korpora von heise-online werden dem K2003 zufällig gewählte Meldungen als Testmenge entnommen und als Soll-Ergebnis manuell ausgezeichnet. Precision und Recall beziehen sich entsprechend auf diese Vergleichsmenge von klassifizerten Begriffsinstanzen. Bei den Korpora von Reuters wird analog vorgegangen.

3.3 Programmierung

Um auf einen Sachverhalt referenzieren zu können benötigt er eine Bezeichnung – so auch die entwickelte Vorgehensweise bzw. das daraus resultierende Programm. Auf Grund des dargestellten Herangehens, das sich der natürlichen Sprache in ihrer geschriebenen Form unterwirft, soll der verfolgte Ansatz sowie die im Folgenden dargestellte Umsetzung ***Seschat*** heißen. Dies ist der Name der ägyptischen Göttin der Schrift und steht für ‚***SEmantic Structures of Communication to enHAnce Textmining***'.

Zur Programmerstellung wurde eine sequenzielle Vorgehensweise in Form eines einzelnen Thread gewählt. Die parallele Abarbeitung einzelner Schritte, beispielsweise während der Auszeichnungsphase, ist auf Grund der gegebenen Hardwarevoraussetzungen nicht sinnvoll. Die wichtigsten Verarbeitungsschritte sind in Abbildung 26 auf Seite 115 dargestellt.

3.3.1 Aufbau der Faktenbasis

Die Datei Seed.xml (siehe CodeTeil 1) beinhaltet die anfänglich gegebenen Beispielinstanzen und entspricht in ihrem Aufbau der Faktenbasis. Für jedes Konzept wird ein eindeutiger Name ‚***name***' sowie ein Attribut ‚***type***' vergeben, welches den morphologischen Typ, also das zu verwendende Mikromuster (siehe S. 93), beschreibt.

CodeTeil 1: Beispiel für 'Seed.xml'

```
<?xml version="1.0" encoding="ISO-8859-1"?>
  <Instanzen>
    <Konzept name="ROLLE" type="11">
      <Bsp>Vorstand</Bsp>
      <Bsp>Chef</Bsp>
      <Bsp>Sprecher</Bsp>
    </Konzept>
    <Konzept name="FIRMA" type="40">
      <Bsp>Deutsche Telekom</Bsp>
      <Bsp>Apple</Bsp>
      <Bsp>Infineon</Bsp>
    </Konzept>
    <Konzept name="PERSON" type="30">
      <Bsp>Ulrich Schumacher</Bsp>
      <Bsp>Steve Jobs</Bsp>
      <Bsp>Heinrich von Pierer</Bsp>
    </Konzept>
    <Konzept name="ORT" type="20">
      <Bsp>Chemnitz</Bsp>
      <Bsp>München</Bsp>
      <Bsp>Leipzig</Bsp>
    </Konzept>
    <Konzept name="BRANCHE" type="11">
      <Bsp>Halbleiter</Bsp>
      <Bsp>IT-Dienstleister</Bsp>
    </Konzept>
  </Instanzen>
```

3.3.1 Aufbau der Faktenbasis

Das Einlesen der XML-Dateien erfolgt mit Hilfe des in Perl enthaltenen Parsers ‚Expat' in einen Hash. Dabei wird für jeden Begriff con_i die Anzahl eingelesener Bezeichner ermittelt und daraus ein Dringlichkeitsfaktor *force* gemäß Formel 12 abgeleitet. Damit wird erreicht, dass Muster zur Erkennung bisher unterdurchschnittlich häufig gefundener Begriffe bevorzugt betrachtet werden (vgl. S.107).

3.3.2 Morphologische Beschreibung

Die mit Hilfe des ‚type'-Attributs verschlüsselten morphologischen Beschreibungen geben das Aussehen der jeweiligen Bezeichner wieder. Eine Möglichkeit zur Abbildung unterschiedlicher Morphologien der Bezeichner eines Begriffs ist die Vergabe mehrerer Typen pro Begriff oder die Anlage mehrerer Begriffe mit je einem Mikromustertypen. In Tabelle 9 sind die Muster am Beispiel des Begriffs <ROLLE> dargestellt.

Tabelle 9: Unterschiedliche Mikromuster eines Begriffs

Typ	01	02
Instanz	Minister der Finanzen	Finanzminister
Mikromuster	[A-Z][a-z]+ (der|des) [A-Z][a-z]+	[A-Z][a-z]+

Die XML-Kodierung mit mehreren morphologischen Typen sähe dann so aus:

```
<Konzept name="ROLLE" type="01" type="02">
  <Bsp>Minister der Finanzen</Bsp>
  <Bsp>Finanzminister</Bsp>
        ...
</Konzept>
```

Die Kodierung mit Hilfe mehrerer Begriffe bringt eine bessere Nachvollziehbarkeit:

```
<Konzept name="ROLLE-lang" type="01">
  <Bsp>Minister der Finanzen</Bsp>
  <Bsp>Staatsminister des Inneren</Bsp>
        ...
</Konzept>
<Konzept name="ROLLE-kurz" type="02">
  <Bsp>Finanzminister</Bsp>
  <Bsp>Innenminister</Bsp>
        ...
</Konzept>
```

Aus diesem Grund sowie der damit verbundenen Vereinfachung bei der Suche nach Programmfehlern wird die zweite Vorgehensweise bevorzugt. Da es sich hier lediglich um eine prototypische Umsetzung handelt, die die prinzipielle Vorgehens-

weise verifizieren soll, wird auf eine logische Verknüpfung der beiden Typen verzichtet, obwohl damit unter Umständen die Ergebnisse schlechter ausfallen.

Im Rahmen der prototypischen Realisierung werden die morphologischen Beschreibungen der einzelnen Typen direkt im Programm definiert und in einem Array of Arrays abgelegt, wie in CodeTeil 2 zu sehen. Um die Anpassbarkeit an andere Sprachen und Zeichensätze zu unterstützen, werden die in Frage kommenden Zeichen vorab definiert, da das verwendete Betriebssystem dies nur mangelhaft unterstützt. An dieser Stelle können auch Ergänzungen der Mikromuster um Beschreibungen für andere Sprachen stattfinden. Bei der Übertragung auf ein tatsächliches Anwendungssystem wäre dieser Teil auszulagern.

CodeTeil 2: Definition der Mikromuster

```
$a_lo = 'a-zäöüß';                          # Zeichensatz Klein
$a_hi = 'A-ZÄÖÜ';                           # Zeichensatz Groß
$a_ch = $a_lo.$a_hi;                        # Gsamtzeichensatz

$firma = "([$a_lo]{1,2}|[0-9]|[$a_hi]{1,3})?[0-9&-+$a_hi][0-9$a_ch]
+([ -][$a_hi][$a_ch]+){0,2}";

$person = "([$a_hi][$a_lo]+[ -]?){1,2}( [$a_hi]\.)?( v[ao]n( de
[nr]?)?)? ([$a_hi][$a_lo]+[-]?){1,2}";

$ort = "[$a_hi][$a_lo]{2,}([ -/][$a_hi][$a_lo]{2,})?";

@klein = (" [$a_lo]{5,}",                   # Einzelwort klein
          "([$a_lo]{5,})( [$a_lo]{5,})?"); # 1/2 Einzelorte klein

@gross = ("[$a_hi][$a_lo]*",                # Einzelwort groß
          "[$a_hi]+[-$a_lo]*",              # Einzelwort/Abkürzung
          "[$a_hi][-$a_lo]*(( des| der| für| und)( [$a_hi][-$a_lo]
*)){1,2}");                                 # Bundesamt für ...

@Mikro = ([@klein],[@gross],[$ort],[$person],[$firma]);
```

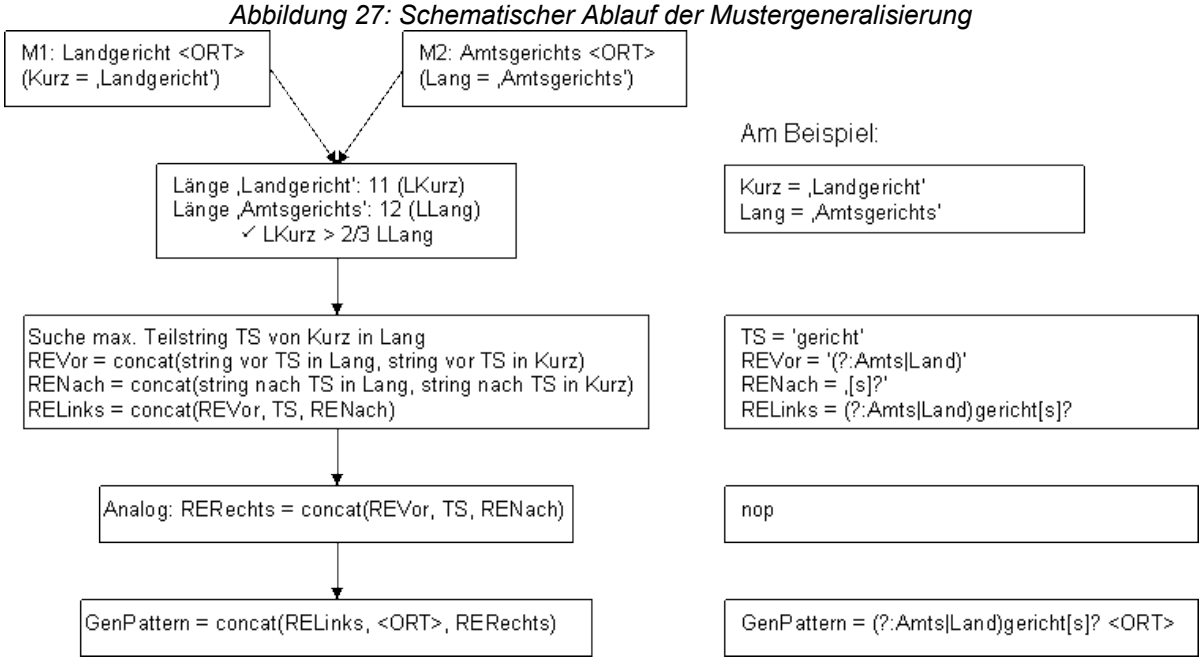
3.3.3 Generalisierung

Die Generalisierung der Muster mit TrustLevel 1 basiert, wie in Abschnitt 3.2.3.5 dargestellt, auf der Suche nach der größten übereinstimmenden Zeichenkette. In Frage kommen Muster des gleichen Begriffs, deren Freitextteile sich um nicht mehr als 1/3 in der Länge unterscheiden. Die Generalisierung wird nacheinander für Zeichenfolgen links und rechts vom Begriff durchgeführt.

3.3.3 Generalisierung

Die Vorgehensweise illustriert Abbildung 27 an Hand der beiden Muster `Landgericht <ORT>` und `Amtsgerichts <ORT>`. Aus diesen wird durch Bildung regulärer Ausdrücke das generalisierte Muster `(?:Amts|Land)gericht[s]? <ORT>`. Damit wird für die Extraktion einerseits nur ein Muster benötigt, was sich positiv auf die Laufzeit auswirkt und andererseits werden damit auch Muster in Betracht gezogen, deren einzelne Signifikanzwerte jeweils zu gering wären.

Abbildung 27: Schematischer Ablauf der Mustergeneralisierung



Eine zweite Generalisierungsstufe betrifft Muster des TL2. Bei diesen wird nach Beendigung eines Lernzyklus versucht, den dazwischen liegenden Textteil durch einen identifizierten Begriff zu ersetzen. Dadurch erhöht sich einerseits der TrustLevel dieses Musters und es lassen sich im darauffolgenden Zyklus weitere Instanzen desselben Begriffs identifizieren. Ein Beispiel dafür ist die Generalisierung von `<ORT>, Dresden, <ORT>`[23] zu `<ORT>, <ORT>, <ORT>`. Die so entstandenen Muster müssen ebenfalls den oben dargestellten Plausibilitätsregeln genügen.

3.3.4 Musteranwendung

Vor der Anwendung der Makromuster werden zu den im aktuellen Zyklus ermittelten Mustern diejenigen mit TrustLevel größer 1 von früheren Läufen aus der Regelbasis hinzu geladen. Die Anwendung erfolgt entsprechend absteigendem TL,

23 Nur bei Aufzählungen dürfen beide Begriffe vom gleichen morphologischen Typ sein.

so dass Muster mit mehr beteiligten Begriffen zuerst betrachtet werden. Für jedes Muster werden nacheinander die enthaltenen Begriffe durch das entsprechende Mikromuster ersetzt und der mit den bereits bekannten Instanzen ausgezeichnete Korpus damit durchsucht. Welches Mikromuster zu wählen ist, wird durch den in der Faktenbasis beim Konzept angegebenen Parameter 'type' gesteuert (siehe CodeTeil 1, S.119).

Das Makromuster `Amtsgericht <ORT>` würde entsprechend des für <ORT> angegebenen Types ‚`20`' zu

```
Amtsgericht (($a_hi][$a_lo]{2,}([ -/][$a_hi][$a_lo]{2,})?)
```

umgeformt. Trifft das so aufbereitete Suchmuster auf die Textstelle ‚`...beim Amtsgericht Düsseldorf eingereicht...`', so bewirkt das in Klammern gesetzte Mikromuster eine Speicherung der Zeichenfolge `Düsseldorf`, die damit als Instanz des Begriffs <ORT> klassifiziert wird.

Vorbereitend zur späteren Bereinigung von Klassifikationsfehlern wird weiterhin abgelegt, wie oft diese Zeichenfolge welchem Begriff zugeordnet wurde und welches Muster dieser Klassifikation zu Grunde lag. Damit lässt sich einerseits beurteilen, ob eine Klassifikation wahrscheinlich richtig ist und andererseits aufgedeckte Fehlklassifikationen sowie deren Folgen beheben.

3.3.5 Fehlerbereinigung

Auf Grund der iterativen Vorgehensweise bei der Klassifikation neuer Instanzen kommt der Erkennung und Bereinigung von Fehlern eine große Bedeutung zu. Insbesondere falsche Klassifikationen von häufig vorkommenden Bezeichnern führen sehr schnell zu unbrauchbaren Ergebnissen.

3.3.5.1 Nachträgliche Musterverletzung

Ein im Deutschen des gegebenen Diskursbereichs sehr häufig auftretendes Kommunikationsmuster ist beispielsweise `<ROLLE> von <FIRMA>`[24]. Allerdings treten auch Formulierungen wie `König von Deutschland` auf, die entsprechend zu einer Klassifikation von `Deutschland` als <FIRMA> führen würden.

24 z.B. aus ‚... wie der Vorstandsvorsitzende von Bosch erläuterte...'

3.3.5 Fehlerbereinigung

Der erste Bereinigungsschritt betrifft die zur Extraktion benutzten Muster der TrustLevel 1 und 2. Enthalten diese Muster eine Zeichenfolge, die im gerade abgeschlossenen Lernschritt als Instanz eines Begriffs identifiziert wurde, so müssen die resultierenden Muster auch den entsprechenden Plausibilitätsregeln wie in Abschnitt 3.2.3.4 dargestellt, entsprechen. Ist dies nicht der Fall sind die durch diese Muster klassifizierten Zeichenfolgen wieder zu deklassifizieren.

CodeTeil 3: Pseudocode zu Deklassifikation

```
M = {erfolgreiche Muster}; I = {klassifizierte Instanzen};

for all m ∈ M
  W = split_to_words(m);
  for all w ∈ W
    if exists (c = concept(w))
    J: for all n ∈ M
      n' = substitute(n,w,c);
      if plausibel(n')
        J: nop;
        N: invalid{n} = w;

for all w,n ∈ invalid
  delete all i ∈ I classified_by(n);
  delete all i ∈ I classifier_of(w);
  delete n; delete classifier_of(w);
```

Die in CodeTeil 3 dargestellte Vorgehensweise sei am folgenden Beispiel illustriert:

Die Textstellen `ein Intel-Deutschland Vertreter sagte` und `wie der Microsoft-Deutschland Chef mitteilte` rufen das Muster `<FIRMA>-Deutschland <ROLLE>` hervor. Wurde auf Grund oben dargestellter Konstellation `Deutschland` als <FIRMA> klassifiziert, so würde das ungültige Muster `<FIRMA>-<FIRMA> <ROLLE>` entstehen. Da nicht entschieden werden kann, welches der beiden Muster den Fehler verursachte, sind alle Instanzen zu deklassifizieren, die durch das Muster `<FIRMA>-Deutschland <ROLLE>` oder `<ROLLE> von <FIRMA>` identifiziert wurden.

3.3.5.2 Mehrfachklassifikation

Mehrfache Klassifikation bedeutet, dass ein Bezeichner unterschiedlichen Klassen zugeordnet wurde. Geschah dies für jede Klasse nur genau einmal, so ist die Richtigkeit der Klassifikation fragwürdig und wird rückgängig gemacht. Ansonsten gilt die Überlegung, dass die vertauenswürdigsten Muster zuerst angewandt worden sind. Damit sind auch deren Ergebnisse vertrauenswürdiger und haben mehr Gewicht. Erst wenn ein Wort mindestens 1,5 mal so oft einer anderen Klasse als der ersten zugeordnet wurde, wird diese alternative Klassifikation als korrekt angenommen. Die Ergebnisse dieser Untersuchungen werden wie in CodeTeil 4 dargestellt protokolliert.

Auslöser für mehrfache Klassifikation sind insbesondere Bezeichner, die ihrerseits einen Begriff im Sinne des Abschnitts 2.1.2, also eine Abstraktionsebene bzw. einen Gattungsbegriff darstellen.

3.3.5.3 Statistische Bereinigung

Mit Hilfe der statistischen Informationen erfolgt eine letzte Bereinigung der Instanzen. Dabei werden diejenigen Zeichenfolgen gelöscht, die sowohl mit kleinem als auch mit großem Anfangsbuchstaben beginnen können, für die aber kein hinreichend hoher TF-IDF-Wert ermittelt werden konnte. Diese Merkmalskombination trifft in erster Linie auf Fehlklassifikationen auf Grund sprachlicher Ambiguitäten oder willkürlicher Wortschöpfungen zu.

CodeTeil 4: Ausschnitt eines Fehlerprotokolls

```
...
mehrfach klassifizierte Instanzen:
Firmen als <BRANCHE> 1 mal
Firmen als <FIRMA> 1 mal  primär:<BRANCHE> unentschieden... löschen
Konzern als <ROLLE> 1 mal
Konzern als <FIRMA> 3 mal  derzeitig:<ROLLE> ändern auf: <FIRMA>
Teste auf Abstraktionsebenen
lösche Oder w/substanti
ändere Nach Elpida auf Elpida
lösche Online w/substanti
ändere Auch PeopleSoft auf PeopleSoft
...
```

3.3.5 Fehlerbereinigung

Bei Instanzen aus mehreren Worten, insbesondere im Zusammenhang mit Firmenbezeichnungen, werden ebenfalls gelegentlich substantivierte Worte als Bestandteil der Instanz betrachtet. Auch diese Fehler lassen sich mit Hilfe der statistischen Daten korrigieren. Einen Ausschnitt eines Fehlerprotokolls zeigt CodeTeil 4.

3.3.6 Ontologische Aspekte

Der Fehlerbereinigung schließt sich die Untersuchung an, wie sich die gefundenen Instanzen innerhalb der gegebenen Begriffswelt einordnen. Dazu wird untersucht, ob eine klassifizierte Zeichenfolge als Namensbestandteil anderer Instanzen auftritt. Da häufig benutzte Worte der natürlichen Sprache im Gebrauch verkürzt werden, wird die kürzere der beiden jeweils betrachteten Instanzen als übergeordnet bewertet. Diese Vorgehensweise erscheint insbesondere vor dem Hintergrund der Generalisierungs-Spezialisierungs-Relationen (siehe S.103) zumindest für den deutschen und schwedischen Korpus berechtigt. Wie die Ergebnisse in Tabelle 6 auf Seite 69 nahe legen, weisen diese Sprachen viele Komposita auf. CodeTeil 5 stellt den Algorithmus in Pseudocode dar. Dabei bezeichnet `cnt` einen Hash, der die Häufigkeit jedes potentiellen Begriffs (`PotBeg`) aufnimmt.

CodeTeil 5: Ermittlung von Unterbegriffen und Concept Links

```
L = F = {klassifizierte Instanzen};
cnt = (PotBeg, Zähler);

for all l ∈ L
  for all f ∈ F
    if (l != f) && (f =~ l)      # Bezeichner ungleich aber ähnlich
    J: cnt{l} += 1;                      # Bezeichner l wird PotBeg
      if begriff(l) == begriff(f)        # gleich klassifiziert
      J: may_Hyponym(l,f);               # eintragen in Hash i.S.v.
      N: may_Link(l,f);                  # l ist Hyponym/Link zu f

for all (PotBeg, Anzahl) ∈ cnt
  if Anzahl > 2
  J: if exists may_Hyponym(PotBeg)        # Hash-Eintrag existiert?
     J: is_Hyponym(PotBeg,begriff(PotBeg));   # PotBeg ist Hypo
       Anzahl --;
  if Anzahl > 2
  J: if exists may_Link(PotBeg)           # PotBeg ist Link
     J: is_Link(begriff(PotBeg),begriff(may_link(PotBeg)));
  delete Instanz{PotBeg};                 # identifizierte Begriffe
                                          # als Bezeichner löschen
```

Ergebnis dieser Untersuchung sind einerseits Begriffe innerhalb eines Oberbegriffs, die das Hyperonym strukturieren und damit die Ontologie stärker detailieren (vgl. Abschnitt 3.2.4, vergleiche auch [FaSt02]). Andererseits treten Begriffe als Verbindung zwischen Begriffen hervor. Diese seien als ***Begriffsbrücken*** oder ***Concept Links*** bezeichnet. Die Ergebnisse einer solchen Analyse sind in CodeTeil 6 dargestellt.

CodeTeil 6: Ergebnisse der ontologischen Analyse

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<OntoData>
  <EBENE name="Hersteller" sup="BRANCHE"></EBENE>
  <EBENE name="Unternehmen" sup="FIRMA"></EBENE>
  <EBENE name="Softwarehersteller" sup="BRANCHE"></EBENE>
  <EBENE name="Chef" sup="ROLLE"></EBENE>
  <EBENE name="Software" sup="BRANCHE"></EBENE>
  <EBENE name="Konzern" sup="FIRMA"></EBENE>
  <LINK name="Hersteller" sup="BRANCHE" to="ROLLE"></LINK>
  <LINK name="Marketing" sup="FIRMA" to="ROLLE"></LINK>
  <LINK name="Konzern" sup="FIRMA" to="ROLLE"></LINK>
</OntoData>
```

Diese beiden Aspekte bieten neben den identifizierten Kommunikations- und damit Begriffsmustern einen wichtigen Ansatzpunkt zum ***Ontology Enrichment*** bzw. ***Ontology Verification*** [Leitn06, S.126]. Allerdings geht dieser tiefe Bezug zur ontologischen Modellierung des Diskursbereichs über die hier zu bearbeitende Thematik bei Weitem hinaus, so dass diese Ergebnisse nur dokumentiert, nicht aber weiter verarbeitet werden. Lediglich die als Unterbegriffe identifizierten Bezeichner der Begriffsbrücken werden gelöscht.

Da wie dargestellt derartige Begriffe oft synonym für konkrete Bezeichnungen von Instanzen gebraucht werden [Biem+03], könnten sie ebenfalls im Sinne einer latenten Klassifikation genutzt werden, etwa um Ambiguitäten aufzulösen.

Aus den in diesem Zusammenhang erzielten Ergebnissen wird einmal mehr deutlich, dass maschinelle Verfahren lediglich unterstützend bei der Erstellung oder Ergänzung bzw. Überprüfung von Ontologien wirken können [Domi+01].

4 Ergebnisse und Diskussion

Den in der Evaluationsplanung dargestellten Schwerpunkten sollen einige allgemeine Aspekte voran gestellt werden, die für die weitere Vorgehensweise wichtig sein können. Dies betrifft die Frage einer geeigneten Lernstrategie, eventuell unter Maßgabe konkreter Einsatzziele, sowie des zu erbringenden zeitlichen Aufwands.

4.1 Vergleichsdaten

Zur Beurteilung der an Hand der Korpora K2003 und K9704 erreichten Ergebnisse wurde die gleiche Testmenge herangezogen. Dabei handelt es sich um 20 dem K2003 vorab entnommene Meldungen. Für diese Texte wurde manuell bestimmt, welche Instanzen der Begriffe erwartet werden. Die erwarteten Werte sind in Tabelle 10 zahlenmäßig dargestellt.

Tabelle 10: Erwartete Instanzen pro Begriff

Begriff	Soll-Instanzen
Branche	12
Firma	38
Ort	16
Person	13
Rolle	21
gesamt	100

Zur Bewertung der hier erreichten Werte werden die Ergebnisse von [CoNLL03] herangezogen, wie sie Tabelle 11 wiedergibt. Diese wurden mit den aus dem Verfahren gewonnenen Ergebnissen maschinell verglichen. Da sich die dargestellten Werte lediglich auf die Named Entities <PERSON>, <FIRMA> und <ORT> beziehen werden bei den Untersuchungen auch die Ergebnisse dieser drei Begriffe explizit betrachtet.

Auffällig bei den Werten in Tabelle 11 ist, dass die Recall-Werte deutlich hinter denen für Precision zurück bleiben. Bestenfalls werden lediglich 66% aller enthaltenen Instanzen gefunden. Damit können Werte für Recall ab 0,60 und Precision ab 0,75 als gutes Ergebnis gewertet werden, was einem Wert von 0,67 im f-Measure, also dem harmonischen Mittel (vgl. S.29) entspricht. Bei den englischen Vergleichsdaten bewegen sich beide Werte gleichmäßig zwischen etwa 0,89 und 0,8. Lediglich vier Gruppen erreichen weniger als 0,8 im f-Measure.

Tabelle 11: Vergleichsergebnisse Deutsch aus [CoNLL03]

Gruppe	precision	recall	F
[FIJZ03]	83.87%	63.71%	72.41±1.3
[KSNM03]	80.38%	65.04%	71.90±1.2
[ZJ03]	82.00%	63.03%	71.27±1.5
[MMP03]	75.97%	64.82%	69.96±1.4
[CMP03b]	75.47%	63.82%	69.15±1.3
[BON03]	74.82%	63.82%	68.88±1.3
[CC03]	75.61%	62.46%	68.41±1.4
[ML03]	75.97%	61.72%	68.11±1.4
[MLP03]	69.37%	66.21%	67.75±1.4
[CMP03a]	77.83%	58.02%	66.48±1.5
[WNC03]	75.20%	59.35%	66.34±1.3
[CN03]	76.83%	57.34%	65.67±1.4
[HV03]	71.15%	56.55%	63.02±1.4
[DD03]	63.93%	51.86%	57.27±1.6
[WP03]	71.05%	44.11%	54.43±1.4
[Ham03]	63.49%	38.25%	47.74±1.5
baseline	31.86%	28.89%	30.30±1.3

4.2 Lernstrategie

Die Umsetzung der Konzeption erfolgte so, dass zwischen einem normalen Lernschritt inklusive der Erschließung neuer Muster (***Learn***) und einem Anwendungsschritt (***Reuse***) gewählt werden kann. Bei letzterem werden lediglich bisher gelernte Fakten und Regeln angewandt, um neue Fakten zu erschließen. Eine Analyse hinsichtlich neuer Muster findet nicht statt (vgl. Abschnitt 3.2.6 auf Seite 114).

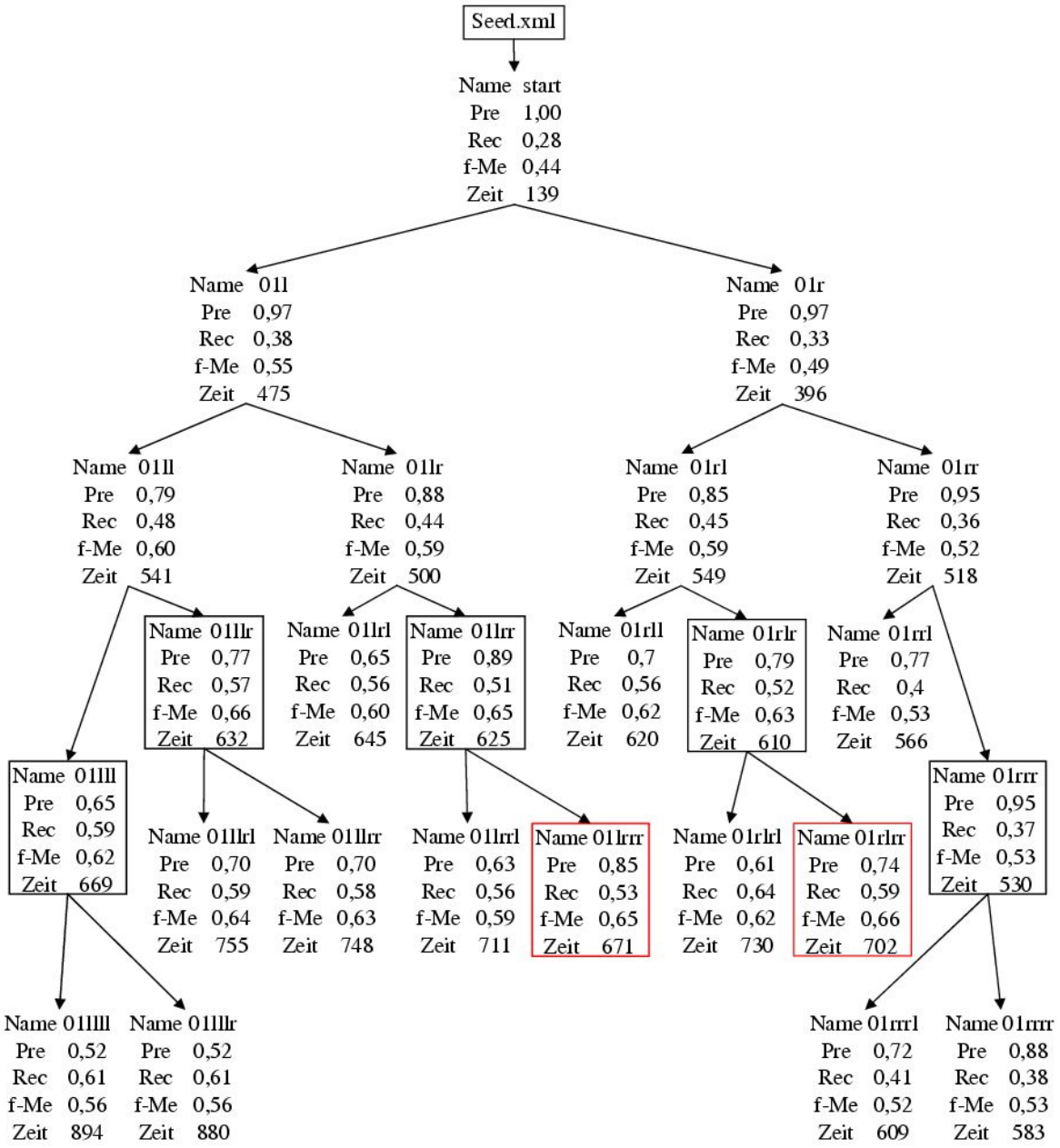
Ziel dieser inkrementellen, steuerbaren Vorgehensweise ist es, ein zwischen Recall und Precision ausgeglichenes Ergebnis zu erreichen bzw. eine der beiden Kenngrößen gezielt zu entwickeln. Gute Werte im Precision sind eher bei ***Reuse*** zu erwarten, während die Erschließung vieler Regeln im ***Learn*** hohe Recall-Werte erwarten lässt. Die Untersuchungen bezüglich einer geeigneten Lernstrategie sind Gegenstand der folgenden Abschnitte. Dabei ist unter einer Iteration ein zwischen ***Learn*** und ***Reuse*** unterscheidender Programmaufruf zu verstehen.

4.2.1 Entwicklungskorpus

Bis einschließlich der 3. Iteration wurden alle Strategien am K2003 untersucht, da erwartet wurde, dass sich ab da nicht Erfolg versprechende Wege ausschließen lassen. Den bearbeiteten Baum zeigt Abbildung 28, wobei Pfeile nach ***links Learn*** und nach ***rechts Reuse*** entsprechen. Entsprechend werden an die Knotennamen

die Buchstaben l bzw. r angehängt. Die Zeiten sind die für den jeweiligen Schritt benötigten Sekunden.

Abbildung 28: Suchbaum einer geeigneten Strategie



Bemerkenswert ist, dass baseline (0,30) im f-Measure schon im ersten Lauf (Name: start) erreicht wird. Bereits nach drei Iterationen bestätigte sich die Erwartung bezüglich Recall- und Precision-Verhalten bei reinen Learn- oder Reuse-Strategien.

Bis zu diesem Punkt sind die Laufzeitunterschiede der verschiedenen Strategien gering, wobei sich kürzere Zeiten für Reuse-Schritte abzeichnen. Die kombinierten Strategien in Iterationsstufe 3 zeigen Werte für Precision im oberen Viertel der Vergleichsdaten. Bezüglich Recall liegen die Ergebnisse im unteren Bereich.

Zur Durchführung der weiteren Schritte wurden die drei Zustände der gemischten Strategien ausgewählt, die den höchsten Wert des f-Measure erreichen. Außerdem wurden die reinen Strategien weitergeführt um zu prüfen, wieweit sich der oben beschriebene Trend fortsetzt. Diese Zustände sind in Abbildung 28 durch schwarze Umrandung hervorgehoben.

Die folgenden Läufe bestätigten die geringe Eignung reiner Strategien. Ausschließliche Anwendung von Learn-Schritten verschlechtert das Ergebnis nach drei Iterationen. Vor dem Hintergrund des zeitlichen Aufwands zur Erreichung des Zustands 01lll ist die Verbesserung so gering, dass dieser Schritt ebenfalls entfallen kann. Im reinen Reuse-Fall ist bereits die zweite Iteration unnötig, da alle nachfolgenden Schritte keine nennenswerte Verbesserung erreichen können.

Die besten Ergebnisse wurden nach drei Iterationen im Zustand 01llr und 01lrr sowie nach vier Iterationen in den Zuständen 01rlrr und 01lrrr erreicht, wobei der Zuwachs im Recall bei letzterem um 0,02 durch einen Rückgang in Precision um 0,04 teuer erkauft wurde. Mithin sind die Strategien zu den Zuständen 01llr sowie 01rlrr die Besten, können die gesteckten Erwartungen jedoch nur zum Teil erfüllen. Vor dem Hintergrund der in der 4. Stufe zusätzlich benötigten 702 Sekunden ist der zum 01llr führenden Strategie der Vorzug zu geben.

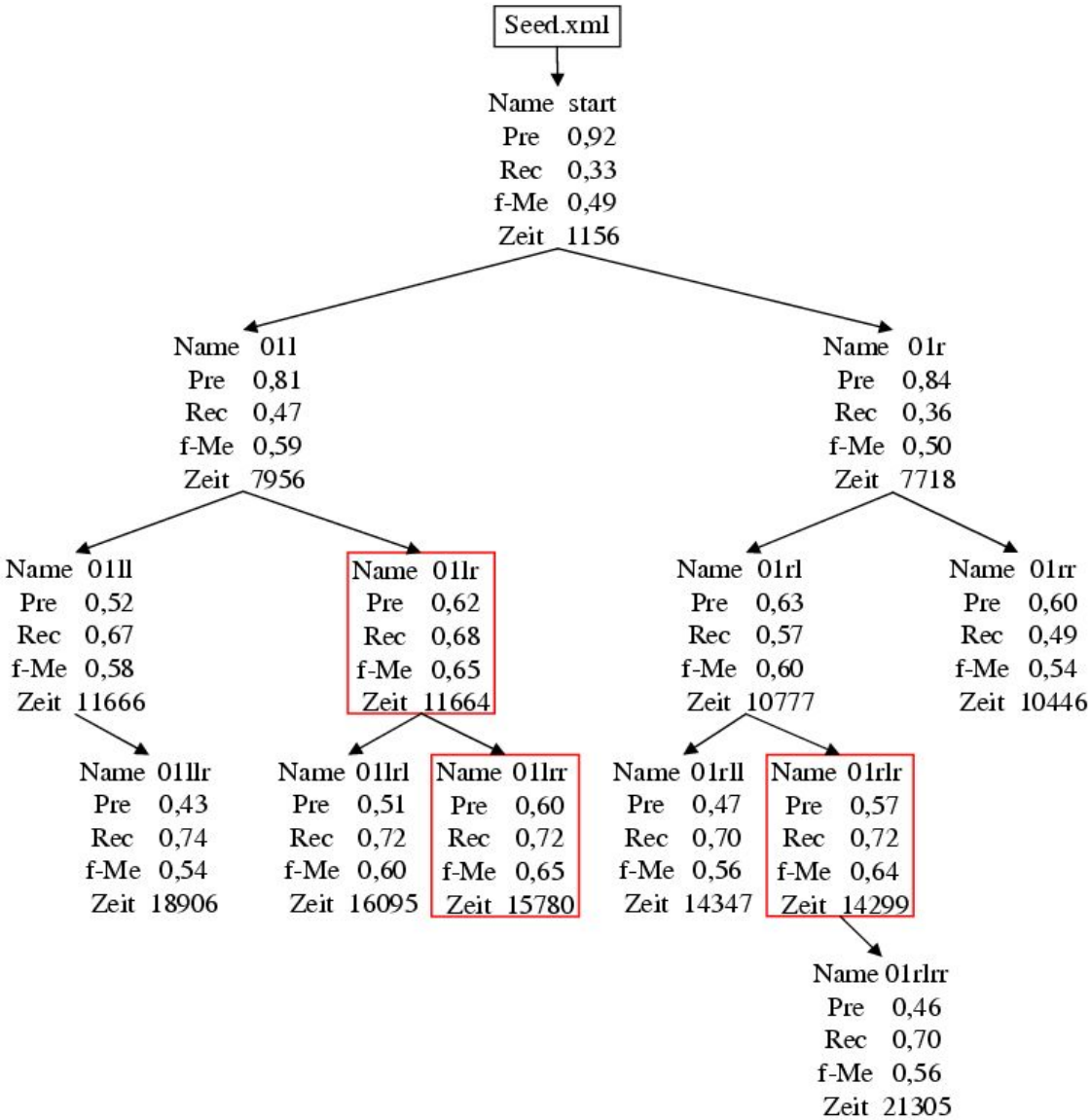
Die Endergebnisse in der 4. Stufe werden in weniger als 45 Minuten erreicht. Obwohl praktisch keine Vergleichsdaten vorliegen, ist es unwahrscheinlich, dass statistische Verfahren ähnlich schnell sind, insbesondere unter Einbeziehung des zur Erstellung der Trainingsmenge nötigen Aufwands.

Die erreichten Resultate werfen die Frage auf, ob insbesondere die Werte des Recall durch eine größere Datenbasis verbessert werden können und wie sich dann die Laufzeit entwickelt.

4.2.2 Skalierung

Zu diesem Zweck sowie zur Überprüfung der oben entwickelten Lernstrategie wird der Korpus K9704 herangezogen, der Meldungen der gleichen Domäne und Quelle, jedoch aus den Jahren 1997 bis 2004 enthält. Es wurden die in Abbildung 29 dargestellten Schritte durchgeführt. Die Läufe zur Erreichung der Zustände 01lll und 01rlrl wurden nach jeweils etwa 16 Stunden abgebrochen, da das zu erwartende Ergebnis in keinem Verhältnis zur Laufzeit stand.

Abbildung 29: Untersuchung des Skalierungsverhaltens



Wie erwartet steigen die Werte im Recall stark an auf bis zu 0,74, was deutlich oberhalb der Vergleichsdaten in Tabelle 11 liegt. Allerdings geht gleichzeitig die Precision zurück, so dass die Werte im f-Measure nicht über die am kleineren Korpus erreichten hinaus gehen. Als ein Grund dafür kommt in Betracht, dass sich innerhalb der 8 Jahre denen die Meldungen entstammen die Kommunikationsmuster geändert haben. Wichtiger dürfte jedoch sein, dass fehlerhafte Klassifikationen auf Grund des häufigeren Auftretens dieser Instanzen einen stärkeren Einfluss auf das Gesamtergebnis haben.

Bezüglich der Lernstrategie lässt sich zusammenfassend feststellen, dass auf den ersten Programmlauf ein weiterer Learn-Vorgang folgen soll. Bei kleinen Korpora wird dem abschließenden Reuse-Schritt ein weiterer Learnzyklus vorangestellt. In Anbetracht der Laufzeiten ist kleinen Korpora der Vorzug zu geben. Der schnellste Weg zur Erreichung des besten f-Measure beträgt zwar auch bei dem großen Korpus nur 20776 Sekunden also etwa 5¾ Stunden. Dennoch liegt das Ergebnis marginal unter dem des kleineren Korpus, bei dem lediglich 1787 Sekunden oder knapp 30 Minuten benötigt wurden.

Bezüglich des Textvolumens kann man den Daten entnehmen, das der verwendete Algorithmus mit etwa O(2n) skaliert. Wegen des geringen Umfangs von lediglich zwei Durchläufen ist diese Angabe jedoch nur bedingt aussagekräftig. Außerdem hängt das Laufzeitverhalten nicht nur vom Volumen des Korpus sondern auch von den gefundenen Regeln sowie den dadurch identifizierten Instanzen ab. Je mehr von beidem bereits gefunden wurde, desto länger dauert die weitere Verarbeitung, wie die Zeitangaben im Lauf der Iterationsschritte zeigen. Hier lässt sich keine allgemeine Aussage treffen.

Den nächsten Schwerpunkt bildet die Untersuchung des Fehlerverhaltens. Dazu werden die bisherigen Ergebnisse auf Begriffsebene betrachtet.

4.3 Fehlerverhalten

Zur Untersuchung des Fehlerverhaltens sollen zunächst die Ergebnisse am Entwicklungskorpus herangezogen werden. Die für die einzelnen Begriffe erzielten Ergebnisse zeigt Tabelle 12. Hier wird deutlich, dass lediglich der Begriff <ROLLE>

unbefriedigende Ergebnisse erzielt. Neben dem geringen Recall von weniger als 0,5 ist insbesondere bei den beiden insgesamt besten Zuständen die Precision zu gering. In der Summe scheitern die Zustände 01lrr sowie 01lrrr in erster Linie an den unbefriedigenden Recall-Werten.

Tabelle 12: Detailergebnisse der besten Läufe mit K2003

Konzept		**01llr**					**01lrr**					**01lrrr**					**01rlrr**				
	Soll	**Ist**	**Fehl**	**Pre**	**Rec**	**Fm**	**Ist**	**Fehl**	**Pre**	**Rec**	**Fm**	**Ist**	**Fehl**	**Pre**	**Rec**	**Fm**	**Ist**	**Fehl**	**Pre**	**Rec**	**Fm**
Branche	12	11	2	0,85	0,92	0,88	11	1	0,92	0,92	0,92	11	2	0,85	0,92	0,88	11	2	0,85	0,92	0,88
Firma	38	23	3	0,88	0,61	0,72	19	3	0,86	0,50	0,63	20	5	0,80	0,53	0,63	22	5	0,81	0,58	0,68
Ort	16	9	1	0,90	0,56	0,69	8	1	0,89	0,50	0,64	8	1	0,89	0,50	0,64	9	1	0,90	0,56	0,69
Person	13	6	0	1,00	0,46	0,63	5	0	1,00	0,38	0,56	6	0	1,00	0,46	0,63	7	0	1,00	0,54	0,70
Rolle	21	8	11	0,42	0,38	0,40	8	1	0,89	0,38	0,53	8	1	0,89	0,38	0,53	10	13	0,43	0,48	0,45
Summen	100	57	17				51	6				53	9				59	21			
Precision	0,75	**0,77**					**0,89**					**0,85**					**0,74**				
Recall	0,60	**0,57**					**0,51**					**0,53**					**0,59**				
f-Measure	0,67	**0,66**					**0,65**					**0,65**					**0,66**				

Betrachtet man die entsprechenden Ergebnisse am K9704, so ergibt sich das in Tabelle 13 dargestellte Bild.

Tabelle 13: Detailergebnisse der besten Läufe mit K9704

Konzept		**01lr**					**01lrr**					**01rlr**				
	Soll	**Ist**	**Fehl**	**Pre**	**Rec**	**Fm**	**Ist**	**Fehl**	**Pre**	**Rec**	**Fm**	**Ist**	**Fehl**	**Pre**	**Rec**	**Fm**
Branche	12	7	4	0,64	0,58	0,61	7	7	0,5	0,58	0,54	7	10	0,41	0,58	0,48
Firma	38	28	12	0,7	0,74	0,72	29	10	0,74	0,76	0,75	30	18	0,63	0,79	0,7
Ort	16	12	3	0,8	0,75	0,77	12	3	0,8	0,75	0,77	11	3	0,79	0,69	0,73
Person	13	9	0	1	0,69	0,82	11	0	1	0,85	0,92	11	1	0,92	0,85	0,88
Rolle	21	12	22	0,35	0,57	0,44	13	28	0,32	0,62	0,42	13	22	0,37	0,62	0,46
Summen	100	68	41				72	48				72	54			
Precision	0,75	**0,62**					**0,6**					**0,57**				
Recall	0,60	**0,68**					**0,72**					**0,72**				
f-Measure	0,67	**0,65**					**0,65**					**0,64**				

Auch in diesem Fall erreicht <ROLLE> die schlechtesten Ergebnisse, wenngleich der Recall deutlich besser geworden ist. Allerdings sind nun die Werte für den Begriff <BRANCHE> in beiden Kategorien schlechter. Zusammenfassend lässt sich sagen, dass insbesondere abstrakte Begriffe von kleinen Korpora profitieren. Die vielen Fehlklassifikationen (Spalten ‚Fehl') zeigen weiterhin, dass diese Begriffe häufig von Klassifikationsfehlern betroffen sind.

Betrachtet man lediglich die auch bei den Vergleichsdaten untersuchten Begriffe, so ist das mit Seschat erreichte Ergebnis sehr gut. Diese Zeilen sind in Tabelle 13 farbig hinterlegt. Insbesondere in den Zuständen 01lr und 01lrr liegen die Ergebnisse im Spitzenbereich. Wie erwartet zeigen sich hier die größten Schwierigkeiten beim

Begriff <FIRMA>, der die komplexeste Morphologie aufweist. Außerdem werden hierfür oft generalisierende Synonyme wie ‚Unternehmen', ‚Firma' oder ‚Konzern' verwendet, was in der Folge ebenfalls zu Fehlklassifikationen führt.

4.4 Domänenabhängigkeit und Seed

Nach der Lernstrategie und dem Skalierungsverhalten wird im Folgenden der Aspekt des Diskursbereichs untersucht. Insbesondere interessiert hier, welchen Einfluss eine thematische Einschränkung des Korpus auf die erreichbaren Ergebnisse hat. Dazu wird der deutschsprachige Reuters-Korpus RDeu herangezogen. Inhaltlich handelt es sich dabei um Nachrichten aus Politik und Wirtschaft, wobei letzteres neben Unternehmensmeldungen auch volkswirtschaftliche und Börsenmeldungen umfasst.

An dieser Stelle bietet es sich an, auch den Einfluss der initialen Beispiele, also des Seeds, zu untersuchen. Dazu wurde der in CodeTeil 7 dargestellte Seed erstellt.

Neben der thematischen Ausweitung wurde hier als weiterer Begriff ‚Organisation', bezeichnet als ***<ORGAN>***, eingeführt. Der Grund dafür ist die Frage, inwieweit Seschat zwischen diesem und dem semantisch sehr ähnlichen Begriff <FIRMA> trennen kann. Die an Hand der beiden Seeds erreichten Ergebnisse zeigt Tabelle 14 im Vergleich, wobei die Spalten Neu0 und Neu0l dem Seed aus CodeTeil 7 entsprechen.

Es ist zu erwarten, dass thematisch unterschiedlich ausgerichtete Beispiele auch zu unterschiedlich fokussierten Begriffsinstanzen führen. Werden beispielsweise als <ROLLE> politische Rollen angeführt, sind kaum wirtschaftliche Bezeichnungen zu erwarten. Aus diesem Grund wäre es nicht sinnvoll, für einen thematisch eingegrenzten Korpus wie K2003 Beispiele aus einer fremden Domäne zu verwenden. Da jedoch oftmals gleiche Bezeichnungen für unterschiedliche Bedeutungen auftreten[25], sind Fehlklassifikationen zu erwarten.

Wegen der Größe des Korpus von 170 MB wird für beide Fälle jeweils nur der initiale sowie ein Learn-Schritt durchgeführt, da die Ergebnisse dann bereits signifikant sein dürften. Die folgenden Schritte würden jeweils mindestens 3 Stunden be-

25 Beispielsweise Regierungschef und Konzernchef

4.4 Domänenabhängigkeit und Seed

CodeTeil 7: Thematisch weniger eingeschränker Seed

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<Instanzen>
<Konzept name="ROLLE" type="10">
<Bsp>Sprecher</Bsp>
<Bsp>Präsident</Bsp>
<Bsp>Chef</Bsp>
</Konzept>
<Konzept name="FIRMA" type="40">
<Bsp>Deutsche Bank</Bsp>
<Bsp>Siemens AG</Bsp>
<Bsp>BASF</Bsp>
</Konzept>
<Konzept name="PERSON" type="30">
<Bsp>Klaus Kinkel</Bsp>
<Bsp>Manfred Kanther</Bsp>
<Bsp>Heinrich von Pierer</Bsp>
</Konzept>
<Konzept name="ORT" type="20">
<Bsp>Frankfurt</Bsp>
<Bsp>Berlin</Bsp>
<Bsp>Dresden</Bsp>
</Konzept>
<Konzept name="ORGAN" type="12">
<Bsp>Regierung</Bsp>
<Bsp>Ministerium</Bsp>
<Bsp>Europäische Union</Bsp>
</Konzept>
<Konzept name="BRANCHE" type="11">
<Bsp>Elektro</Bsp>
<Bsp>Betreiber</Bsp>
<Bsp>Chemie</Bsp>
</Konzept>
</Instanzen>
```

anspruchen, ohne dass die Ergebnisse diesen Aufwand rechtfertigten. In Tabelle 14 sind die erreichten Ergebnisse dargestellt.

Tabelle 14: Wirkung unterschiedlicher Seeds

Konzept		**Neu0**					**Neu0l**					**Alt0**					**Alt0l**				
	Soll	**Ist**	**Fehl**	**Pre**	**Rec**	**Fm**	**Ist**	**Fehl**	**Pre**	**Rec**	**Fm**	**Ist**	**Fehl**	**Pre**	**Rec**	**Fm**	**Ist**	**Fehl**	**Pre**	**Rec**	**Fm**
Branche	10	3	0	1,00	0,30	0,46	4	0	1,00	0,40	0,57	0	0	0,00	0,00	0,00	0	0	0,00	0,00	0,00
Firma	15	0	0	0,00	0,00	0,00	3	9	0,25	0,20	0,22	0	1	0,00	0,00	0,00	3	10	0,23	0,20	0,21
Organ	17	3	1	0,75	0,18	0,29	3	1	0,75	0,18	0,29	0	0	0,00	0,00	0,00	0	0	0,00	0,00	0,00
Ort	50	26	1	0,96	0,52	0,68	32	6	0,84	0,64	0,73	30	3	0,91	0,60	0,72	33	10	0,77	0,66	0,71
Person	30	2	0	1,00	0,07	0,13	5	0	1,00	0,17	0,29	0	0	0,00	0,00	0,00	1	0	1,00	0,03	0,06
Rolle	18	3	1	0,75	0,17	0,27	3	0	1,00	0,17	0,29	3	1	0,75	0,17	0,27	3	3	0,50	0,17	0,25
Summen	140	37	3				50	16				33	5				40	23			
Precision		**0,93**					**0,76**					**0,87**					**0,63**				
Recall		**0,26**					**0,36**					**0,24**					**0,29**				
f-Measure		**0,41**					**0,49**					**0,37**					**0,39**				

Wie erwartet liegen die Ergebnisse mit angepasstem Seed vor denen mit den ursprünglichen Beispielen. Allerdings ist festzustellen, dass eine fehlende Eingrenzung der Domäne insgesamt zu unbefriedigenden Ergebnissen führt. Lediglich die Werte für den Begriff <ORT> sind in beiden Fällen sehr gut. Der Grund dafür dürfte in der großen Stabilität des entsprechenden Kommunikationsmusters liegen.

Auffällig ist, dass die Trennung zwischen <FIRMA> und <ORGAN> offenbar nicht stattfand. Betrachtet man die Ergebnisse im Detail, wie in CodeTeil 8 dargestellt, zeigt sich, dass viele Organisationen als Firma klassifiziert wurden.

CodeTeil 8: Ergebnisse für <FIRMA> im Detail

```
<FIRMA>: Soll:15; Okay:3; Add:9; prec:0.25; rec:0.20 FM:0.22
fehlend: Xinhua, Asko, Canal Plus, Thomson Multimedia, Universal
Pictures,  Bloomberg, Thomson, DSBK, Alcatel Alsthom, ENI,
Lagardere, Kaufhof,
falsch: HBV, Gesamtmetall, DIHT, Separatisten, SPD, ÖTV, UNO, DGB,
EU,
```

Der Grund dafür ist, dass beide Begriffe auf Grund ihrer semantischen Nähe auch sehr ähnlich kommuniziert werden. Das verdeutlicht das Muster

`<ROLLE> der <ORT>er <FIRMA>`, das die Bezeichner `Behörde, Staatsanwaltschaft, Polizei ,Firma RMS Titanic, Kantonspolizei, Opposition, Messe Manfred Wutzhofer, Bundesanstalt, Textilholding Dierig AG, Feuerwehr, Rückversicherungsgesellschaft AG, Vertretung, Rück, Konzertagentur Mama Concerts` und `SPD-Fraktion` hervorbrachte. Da diese Bezeichnungen in der Testmenge nicht vorkommen, treten sie in dem in CodeTeil 8 dargestellten Protokoll nicht als Fehlklassifikation auf.

Zur eigentlichen Untersuchung der Abhängigkeit von der Eingrenzung der Domäne werden dem Gesamtkorpus 15.000 Meldungen zufällig entnommen, so dass sich ein Volumen von etwa 22 MByte ergibt. Die Begriffe <ORGAN> und <FIRMA> wurden zusammenfassend als Organisation betrachtet. Die beiden Zustände in denen die besten Ergebnisse erreicht wurden, sind in Tabelle 15 dargestellt.

Die Ergebnisse beziehen sich wiederum auf eine daraus vorab entnommenen Menge von Meldungen, die manuell ausgezeichnet wurden.

Tabelle 15: Beste Ergebnisse ohne Eingrenzung am RDeu

Konzept		**01ll**					**01llr**				
	Soll	**Ist**	**Fehl**	**Pre**	**Rec**	**Fm**	**Ist**	**Fehl**	**Pre**	**Rec**	**Fm**
Branche	9	3	0	1,00	0,33	0,50	3	0	1,00	0,33	0,50
Organisation	53	25	4	0,86	0,47	0,61	32	4	0,89	0,60	0,72
Ort	50	23	12	0,66	0,46	0,54	27	15	0,64	0,54	0,59
Person	33	10	0	1,00	0,30	0,47	12	0	1,00	0,36	0,53
Rolle	19	8	15	0,35	0,42	0,38	11	27	0,29	0,58	0,39
Summen	164	69	31				85	46			
Precision		**0,69**					**0,65**				
Recall		**0,42**					**0,52**				
f-Measure		**0,52**					**0,58**				

Wie bereits am K2003 (siehe S. 130) bring auch hier der Zustand 01llr das beste Resultat, so dass sich die oben dargestellte Lernstrategie bestätigt. Die Ergebnisse bleiben deutlich hinter denen bei eingeschränktem Diskursbereich zurück. Betrachtet man den dargestellten letzten Schritt von 01ll nach 01llr, steht einer starken Zunahme des Recalls eine eher geringe Verschlechterung der Precision gegenüber, was vor allem der hohen Zahl zusätzlicher Organisationen zuzuschreiben ist.

Auffällig ist an dieser Stelle das schlechte Ergebnis für den Begriff <ROLLE>. Dieses erklärt sich beim Blick auf das in CodeTeil 9 wiedergegebene Testprotokoll. Die unterstrichenen Instanzen stellen Bezeichnungen dar, die in Abhängigkeit der Pragmatik Gruppen als Träger von Rollen oder als Entitäten bezeichnen.

CodeTeil 9: Bewertungsprotokoll Zustand 01llr für den Begriff <ROLLE>

```
<ROLLE>: Soll:19; Okay:11; Add:27; prec:0.29 rec:0.58 FM:0.39
fehlend: Minister, Sicherheitsberater, Vorsitzende, Staatsminister,
Senatskanzleichef, Bundespräsident, Regierungssprecher,
Finanzminister,
falsch: Parteitag, Aktie, Beschäftigte, Spitze, Verpflichtung,
Tochter, Schätzung, Präsidentsschaftswahlen, Politiker, Kritik,
Abgeordnetenhauses, Angaben, Anteil, Pläne, Einschätzung,
Entscheidung, Niederlassung, Abschluß, Mitglieder, Bericht,
Darstellung, Expertengruppe, Ansicht, Betriebsräte, Studie,
Ostdeutschland, Umfrage,
```

Die unterstrichenen, als Fehlklassifikation gewerteten Bezeichner, sind je nach Pragmatik als Gruppe im Sinne einer Organisation oder als Gruppe im Sinne ihrer Funktion zu betrachten.

Zusammenfassend ist festzustellen, dass ein eingeschränkter Diskusrsbereich vorteilhaft für die Anwendung des Verfahrens ist. Die besseren Ergebnisse an thematisch eingeschränkten Korpora scheinen also typisch für extrahierende Verfahren zu sein [FaSt02]. Die Protokolle der aufeinander folgenden Lernschritte zeigen, dass geringe Klassifikationsfehler zu Beginn der Verarbeitung sich noch stärker ausbreiten, da diese Instanzen mit größerer Wahrscheinlichkeit in anderen Konstellationen ebenfalls auftreten. Ein Beispiel dafür ist das Muster ‚`am Donnerstag in <ORT>`'. Damit wurden 111 Instanzen korrekt und 4, darunter ‚`Dienst`', falsch identifiziert. Dies hatte das Muster ‚`<ROLLE> <PERSON> am <ORT>ag in <ORT>`' zur Folge, was weitere Wochentage als Orte klassifizierte.

4.5 Sprachliche Neutralität

Der Schwerpunkt dieser Untersuchungen liegt vor allem in der Fragestellung, ob der dargestellte Ansatz tatsächlich sprachunabhängig ist und welcher Aufwand für eine Anpassung an andere Sprachen zu leisten ist. Daher findet eine thematische Einschränkung an Hand der in der Reuters-Korpora vergebenen Topic-Codes auf Unternehmensmeldungen statt.

4.5.1 Schwedischer Korpus

Der schwedische Korpus wurde auf Grund einiger Besonderheiten zur Überprüfung der sprachlichen Neutralität des Ansatzes herangezogen. Ein Argument ist, dass dazu im Gegensatz zum Englischen kaum Untersuchungen existieren und daher ein gewisser Neuigkeitswert zu erwarten ist. Weiterhin ist der thematisch begrenzte Korpus mit knapp 6MByte sehr klein, so dass statistische Verfahren kaum anwendbar sind. Den für die Untersuchungen erstellten Seed zeigt CodeTeil 10.

Die Anpassung des Zeichensatzes an den schwedischen Korpus erfolge durch Einfügung der beiden Sonderzeichen Å und å in den zu Beginn von CodeTeil 2 auf Seite 121 dargestellten Zeichenvorrat. Diese beiden zusätzlichen Zeichen sowie der deutlich ausgeprägtere Gebrauch von Sonderzeichen gegenüber z.B. dem Deutschen führt dazu, dass built-in-functions zur Verarbeitung und Operatoren zum Vergleich von Zeichenketten wesentlich häufiger versagen.

4.5.1 Schwedischer Korpus

Auffällig ist beim schwedischen Korpus, dass die einzelnen Meldungen sehr kurz und prägnant sind. Das hat zur Folge, dass sich die Texte stets nur einem sehr eingeschränkten Diskursbereich widmen und die thematische Eingrenzung des Korpus sehr gut funktioniert.

Ebenso wie im Englischen werden auch hier nur Eigennamen groß geschrieben. Allerdings werden in Abhängigkeit von der Pragmatik selten auch abstrakte Bezeichnungen wie Eigennamen behandelt. Weiterhin treten auch Abkürzungen abstrakter Begriffe durch Großbuchstaben auf. Daher ist eine zusätzliche morphologische Beschreibung zu erstellen, die dies zulässt.

CodeTeil 10: Seed für den schwedischen Korpus

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<Instanzen>
  <Konzept name="ROLLE" type="00">
    <Bsp>analytiker</Bsp>
    <Bsp>mäklar</Bsp>
    <Bsp>chef</Bsp>
  </Konzept>
  <Konzept name="FIRMA" type="40">
    <Bsp>Scania</Bsp>
    <Bsp>Skanska</Bsp>
    <Bsp>Svenska Bostäder</Bsp>
  </Konzept>
  <Konzept name="PERSON" type="30">
    <Bsp>Göran Ahlström</Bsp>
    <Bsp>Ulf Thorne</Bsp>
    <Bsp>Erik Åsbrink</Bsp>
  </Konzept>
  <Konzept name="ORT" type="20">
    <Bsp>Malmö</Bsp>
    <Bsp>Sundsvall</Bsp>
    <Bsp>Stockholm</Bsp>
  </Konzept>
  <Konzept name="BRANCHE" type="00">
    <Bsp>fastighet</Bsp>
    <Bsp>investmentbank</Bsp>
    <Bsp>försäljning</Bsp>
  </Konzept>
</Instanzen>
```

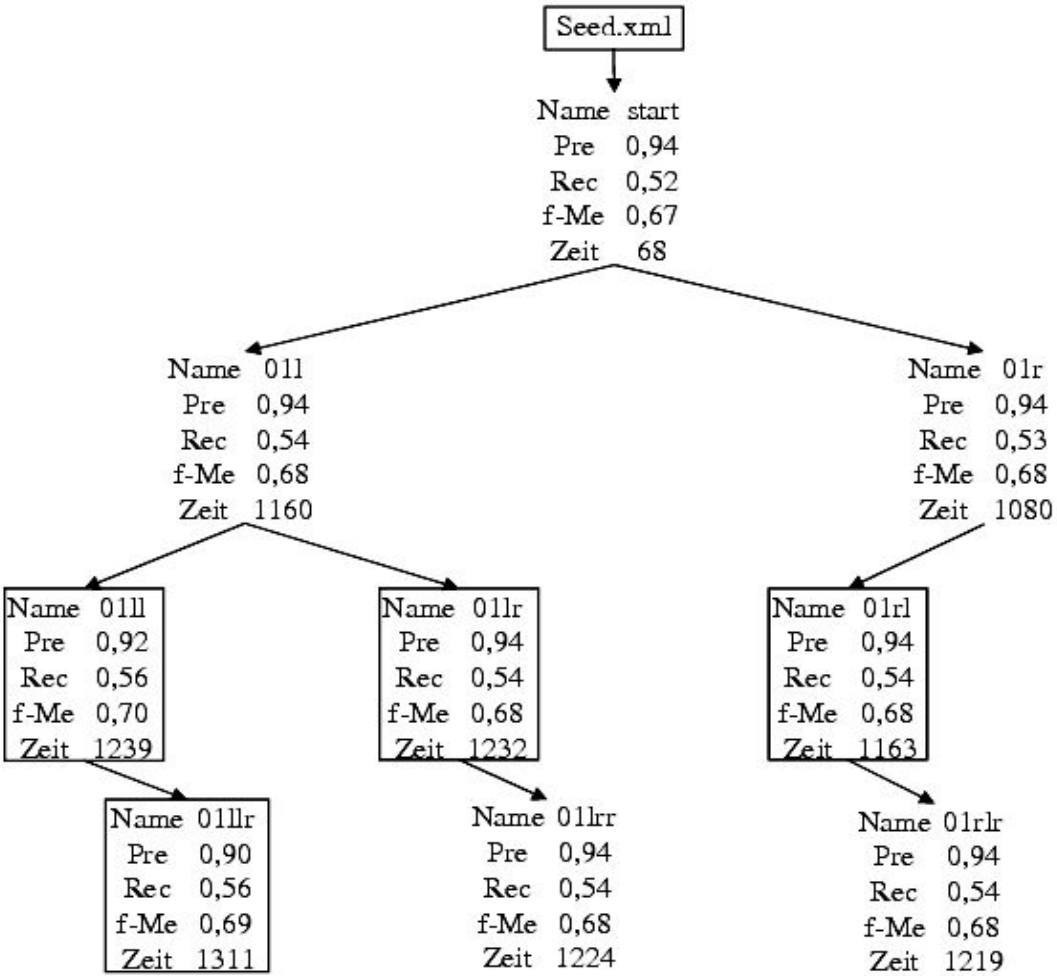
Die Begriffe <BRANCHE> und <ROLLE> sind jeweils als einzelnes Wort ohne Großbuchstaben definiert. Sie bieten somit keine morphologische Unterscheidbarkeit zu beispielsweise Verben oder Adjektiven, weswegen hier der Qualität und Aussagekraft der Extraktionsmuster noch größere Bedeutung zukommt. Durch die

Eingrenzung der Domäne sollten jedoch wenige Ambiguitäten auftreten und diese Problematik nicht zu stark werden.

Bei der Erstellung der Testbasis war auffällig, dass die beiden hier betrachteten abstrakten Begriffe nur sehr selten auftraten. Obwohl dies nicht Teil der thematischen Einschränkung war, traten in erster Linie Meldungen über sich ändernde Besitzverhältnisse an Firmen auf. Personalien hingegen wurden kaum veröffentlicht.

Die Ergebnisse am schwedischen Korpus gibt Abbildung 30 wieder. Auf Grund der geringen Größe und der starken thematischen Einschränkung erfolgt hier wieder die Darstellung als Baum um die Entwicklung besser zu verdeutlichen.

Abbildung 30: Untersuchungen am schwedischen Korpus



Die Änderung der zu Grunde liegenden Sprache hat offenbar keine negativen Auswirkungen. Viel deutlicher wird, dass die eng begrenzte Domäne zu einer sehr steilen Lernkurve führt. Bereits nach dem initialen Lauf wird ein Recall-Wert von 0,52

erreicht, was für eine große Anzahl Instanzen spricht. Entsprechend benötigen die nachfolgenden Schritte selbst im Vergleich zum K2003 eine deutlich längere Laufzeit, verbessern das Ergebnis aber nur geringfügig.

Bis zum Zustand 01II benötigt Seschat knapp 42 Minuten. Ein probehalber durchgeführter Lauf zum Zustand 01III bestätigte die Erwartung, dass dann deutlich schlechtere Ergebnisse erreicht werden. Die Ergebnisse im Detail für die beiden besten Zustände zeigt Tabelle 16.

Tabelle 16: Detailergebnisse für RSve

Konzept		**01II**					**01IIr**				
	Soll	**Ist**	**Fehl**	**Pre**	**Rec**	**Fm**	**Ist**	**Fehl**	**Pre**	**Rec**	**Fm**
Branche	5	4	1	0,80	0,80	0,80	4	1	0,80	0,80	0,80
Firma	57	41	0	1,00	0,72	0,84	41	0	1,00	0,72	0,84
Ort	16	3	0	1,00	0,19	0,32	3	0	1,00	0,19	0,32
Person	14	4	0	1,00	0,29	0,44	4	0	1,00	0,29	0,44
Rolle	6	3	4	0,43	0,50	0,46	3	5	0,38	0,50	0,43
Summen	98	55	5				55	6			
Precision		**0,92**					**0,90**				
Recall		**0,56**					**0,56**				
f-Measure		**0,7**					**0,69**				

Hier wird am Begriff <ORT> deutlich, dass ein kleiner Korpus unter Umständen zu geringen Recall-Werten führt, wenn der betrachtete Begriff in vielen verschiedenen Konstellationen auftritt beziehungsweise in der als typisch erkannten nur selten. Bis zum Zustand 01II wurden insgesamt lediglich 27 Instanzen für <ORT> identifiziert. Im thematisch nicht eingeschränkten Korpus waren dies bis dahin bereits 349.

4.5.2 Englischer Korpus

Im Gegensatz zu den beiden vorher behandelten Sprachen kennt das Englische keine Sonderzeichen. Auch hier wurden 15.000 Meldungen zufällig zu einem thematisch eingeschränkten Teilkorpus zusammengefasst.

Die Meldungen dieses Korpus sind deutlich länger als die der beiden anderen Sprachen und enthalten viele ergänzende Informationen. Im Gegensatz zu den anderen Sprachen findet sich hier in praktisch jeder Meldung mindestens einmal der vollständige Name einer Organisation, der in der weiteren Verwendung jedoch abgekürzt wird. Auf Grund der historischen Entwicklung sowie der amerikanischen Kultur treten oft sehr lange Bezeichnungen wie beispielsweise ‚Muriel Siebert Capital

Markets Group Inc.‘ auf. Diese werden jedoch nicht in ihrer vollen Länge berücksichtigt, weil dies der Allgemeingültigkeit der entsprechenden Morphologie widersprechen würde.

Weiterhin ist auffällig, dass häufiger als im Schwedischen auch abstrakte Begriffe mit großen Buchstaben beginnen. In der zur Verfügung stehenden Zeit konnten jedoch keine Untersuchungen zu diesbezüglichen Regelmäßigkeiten durchgeführt werden. Auch in diesem Korpus finden sich nur sehr wenige Instanzen für <ROLLE> und <BRANCHE>. Es ist anzunehmen, dass die Zielgruppe dieser Meldungen dieser Informationen nicht bedarf, da die Nachrichten vorselektiert werden und entsprechend nur vertraute Bereiche enthalten.

Die einzige zu erwartende technische Schwierigkeit beruht auf der bereits in Abschnitt 2.2.5.1 ab Seite 45 dargestellten Besonderheit, dass im Englischen praktisch keine Komposita auftreten. Insbesondere abstrakte Begriffe sind daher durch zwei und mehr Worte bezeichnet. Entsprechend wurde eine weitere morphologische Beschreibung erstellt, die wie auch im Schwedischen großbuchstabige Abkürzungen wie CEO für chief executive officer erlaubt. In diesem Zusammenhang ist weiter auffällig, dass die in Sprachen mit Komposita auftretenden Generalisierungs-Spezialisierungs-Relationen hier nicht mehr unmittelbar zusammen hängen. Dieser Umstand wurde bei der Programmierung nicht in Betracht gezogen, so dass für diese Fälle nur unbefriedigende Ergebnisse zu erwarten sind.

Auf Grund der bisher als erfolgreich ermittelten Lernstrategie sind in Tabelle 17 nur die Schritte bis zum Zustand 01llr dargestellt. Wegen der schlechten Ergebnisse bei dieser Strategie wurde versucht, über Reuse-Schritte die Precision-Werte zu verbessern, was jedoch nicht gelang.

Tabelle 17: Ergebnisse aus REng

	Soll	Ist	Fehl	Pre	Rec	Fm	Ist	Fehl	Pre	Rec	Fm	Ist	Fehl	Pre	Rec	Fm	Ist	Fehl	Pre	Rec	Fm
Branche	9	1	0	1,00	0,11	0,20	1	0	1,00	0,11	0,20	1	0	1,00	0,11	0,20	1	9	0,10	0,11	0,11
Firma	28	3	0	1,00	0,11	0,19	3	0	1,00	0,11	0,19	4	3	0,57	0,14	0,23	6	3	0,67	0,21	0,32
Ort	29	9	0	1,00	0,31	0,47	11	19	0,37	0,38	0,37	11	19	0,37	0,38	0,37	12	19	0,39	0,41	0,40
Person	22	0	0	0,00	0,00	0,00	0	0	0,00	0,00	0,00	0	0	0,00	0,00	0,00	0	0	0,00	0,00	0,00
Rolle	9	2	0	1,00	0,22	0,36	2	0	1,00	0,22	0,36	3	0	1,00	0,33	0,50	4	14	0,22	0,44	0,30
Summen	97	15	0				17	19				19	22				23	45			
Precision		**1**					**0,47**					**0,46**					**0,34**				
Recall		**0,15**					**0,18**					**0,2**					**0,24**				
f-Measure		**0,27**					**0,26**					**0,28**					**0,28**				

4.5.2 Englischer Korpus

Aus den in Tabelle 17 dargestellten Ergebnissen wird deutlich, dass bei der Anwendung von Seschat auf den englischen Korpus größerer Anpassungsaufwand zu betreiben ist. Als Ursache ist zu sehen, dass die Eigenschaft der Sprache, Begriffe durch eine Vielzahl von Worten zu bezeichnen die Ursache dafür ist. Diese Eigenschaft hat zur Folge, dass die Segmentierung, also die Feststellung welche Zeichenfolgen noch zu einem Bezeichner gehören, erschwert wird. An dieser Stelle kommen die bei einer syntaktischen Analyse ermittelten Wortarten als ergänzende Informationen besonders zum tragen.

Bei der Konzeption wurde offenbar vor dem Hintergrund der im primären Interesse stehenden Sprache Deutsch die Allgemeingültigkeit vernachlässigt. Auch im Englischen finden sich die erwähnten GSR sehr häufig bei abstrakten Begriffen, jedoch treten sie hier nicht in unmittelbarer Verbindung auf. Auf Grund des bei Seschat verfolgten Ansatzes der Begriffe im Zusammenhang verwundert es dann nicht weiter, dass auch die Ergebnisse für die anderen Begriffe unbefriedigend ausfallen.

5 Ausblick und Schluss

Gegenstand der vorliegenden Arbeit war die Fragestellung, ob ein begriffsbasierter Ansatz zur inhaltlichen Erschließung von Texten in natürlicher Sprache realisierbar ist, wie die entsprechende Vorgehensweise aussieht und welche Ergebnisse sich damit erzielen lassen.

Dazu wurde das zu Grunde liegende Verständnis davon, was ein Begriff ist, hinterfragt. Untermauert von Überlegungen zu Kommunikation und Sprachverständnis wurde eine an der Verwendung orientierte Auffassung vom Wesen des Begriff als geeignet identifiziert. Dadurch eröffnete sich die Möglichkeit einer iterativen Vorgehensweise, die im Gegensatz zu statistischen, wortorientierten Verfahren keiner umfangreichen Trainingsdaten bedarf. Im Rahmen der Darstellung aktueller Methoden im Information Retrieval trat immer wieder die grundlegende Bedeutung des Begriffsverständnisses zu Tage.

Als wesentlicher Punkt wurde die Erkennung von Begriffen bzw. deren Bezeichnern in natürlichsprachigen Texten identifiziert. Auf Grund der begriffsorientierten Herangehensweise sollte die Erkennung unabhängig von den auftretenden Bezeichnern und damit der verwendeten Sprache sein. Um den Aspekt der sprachlichen Neutralität zu unterstützen verbot sich die Verwendung von Verfahren zur syntaktischen Analyse.

Basis dieser Erkennung der Begriffsbezeichner sind Kommunikationsmuster, deren maschinelle Erschließung und Anwendung Gegenstand der prototypischen Realisation war. Als Anwendungsbereich wurden wirtschaftlich relevante Nachrichten und als zu Grunde liegende Sprache Deutsch gewählt. Diese Wahl wurde wegen der wirtschaftlichen Bedeutung einerseits und den bestehenden Defiziten bei der Erschließung von Eigennamen im Deutschen andererseits getroffen. Damit ordnet sich diese Arbeit in das Spannungsfeld von wirtschaftlichem Interesse und Informatik ein, wobei hier durch die Basis ‚Sprache' eine weitere Schnittstelle hinzu kommt.

Die gezeigten Ergebnisse sind ein Beleg für die Richtigkeit der vorgeschlagenen Vorgehensweise, die sich am Zusammenhang beziehungsweise der Verwendung von Begriffen orientiert.

Es wurde gezeigt, dass mit Hilfe dieses Ansatzes sowohl typische Eigennamen als auch abstrakte Begriffe erschlossen werden können. Die dabei verfolgte inkrementelle Vorgehensweise gestattet eine deutliche Verringerung des manuellen Aufwands, da an Hand weniger Beispiele gelernt wird. Durch die maschinelle Erschließung von Kommunikationsmustern kann außerdem auf vorab definierte Regeln verzichtet werden.

Da bei Betrachtung von Begriffen die verwendeten Worte sekundär werden, kann weiterhin bei dieser Vorgehensweise auf ein vorheriges Part-of-Speech-Tagging verzichtet werden, was einen wichtigen Schritt in Richtung sprachlicher Neutralität darstellt. Dieser Vorteil zeigte sich auch bei der Anwendung des Verfahrens auf den schwedischen Korpus.

Allerdings traten auch Potentiale für Verbesserungen zu Tage. Ein Nachteil der inkrementellen Vorgehensweise ist, dass sich Fehler sehr stark fortpflanzen können und so das Gesamtergebnis unter Umständen verschlechtern.

Die Fehler rühren einerseits aus der sprachlichen Ambiguität her, was zu echten Fehlklassifikationen führt. Andererseits führen unterschiedliche Abstraktionsstufen innerhalb eines Begriffs ebenfalls zu Fehlern, wobei hier die Konsequenzen größer sind, weil Bezeichner höherer Abstraktionsstufen häufiger und in unterschiedlichen Zusammenhängen vorkommen.

Um diese Probleme zu lösen stellt die begonnene Untersuchung der Bezeichner beziehungsweise der dadurch bezeichneten Begriffe hinsichtlich ihrer ontologischen Bedeutung einen ersten Schritt dar. Hier liegen neben der Fehlervermeidung die größten Möglichkeiten für Erweiterungen, die unter der Bezeichnung ***Ontology Engineering*** zusammengefasst werden können.

Aus Sicht der Wirtschaftsinformatik ist darauf hinzuweisen, dass die vorgestellte Herangehensweise nicht in jedem Fall geeignet ist. Die Mehrzahl der Menschen hat sich auf bestehende Systeme und die dort implementierten Verfahren des Text Retrieval eingestellt und kommt damit sehr gut zurecht. Wenn es jedoch darum geht, aus einer großen Fülle von Informationen gezielt auf begriffliche Einheiten zuzugreifen, dann bietet der dargestellte Ansatz eine sinnvolle Ergänzung. Solche Einheiten betreffen beispielsweise Fragen wie ‚Welche Person war von wann bis

wann Vorstand der Telekom?‘. Derartige Fragestellungen setzen voraus, dass in den zu durchsuchenden Daten die entsprechenden Begriffe wie ‚Person‘ und ‚Rolle‘ identifiziert werden können.

Die vorgestellte Methodik der Informationsextraktion ermöglicht des Weiteren eine Strukturierung und damit Eingrenzung des Suchraums innerhalb anderer Systeme des Information Retrieval, was zu geringerem zeitlichen Aufwand bei der Informationsbeschaffung und damit wirtschaftlichen Vorteilen führt.

Der aufgedeckte Fehler in der Umsetzung bezüglich der Generalisierungs-Spezialisierungs-Relationen in Sprachen ohne Komposita ist ärgerlich. Allerdings ist dies eher als technisches Problem und Herausforderung zu sehen, da es sich hier lediglich um eine prototypische Umsetzung handelte. Die guten Ergebnisse in anderen Fällen belegen die prinzipielle Richtigkeit der auf den Zusammenhang von Begriffen als Basis sprachlicher Kommunikation beruhenden Herangehensweise.

Neben dem zuvor genannten Punkt sind weitere Untersuchungen insbesondere im Bezug auf weitere Arten von Begriffen wie beispielsweise Vorgängen nötig. Ein weiterer Punkt ist die Einbindung der Ergebnisse in die verwendete Ontologie sowie Untersuchungen zur sprachübergreifenden Anwendung. Ein Aspekt in diesem Zusammenhang sind automatischen Übersetzungen, die beispielsweise eine Zuordnung von Kommunikationsmustern unterschiedlicher Sprachen erforderten.

Literaturverzeichnis

[AaPl94] *Aamodt, A.; Plaza, E.*: Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. In: AI Communications, Bd.1 H.7/1994, S.39-59.

[AgYu98] *Aggarwal, C.C.; Yu, P.S.*: Mining Large Itemsets for Association Rules. In: Bulletin of the TC on Data Engineering, Bd.21 H.1/1998, S.23-31.

[AlLe01] *Alavi, M.; Leidner, D.E.*: Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. In: MIS Quarterly, Bd.25 H.1/2001, S.107-136.

[Arist07] *Aristoteles*: Metaphysik - Übersetzung von Adolf Lasson. Jena, 1907.

[AuWe02] *Auth, G.; Wegener, H.*: Die Swiss Re Data Language - Erfahrungen mit Terminologiemanagement im Rahmen des Data Warehousing. In: Vom Data Warehouse zum Corporate Knowledge Center. Heidelberg, 2002, S.193-214.

[Baco20] *Bacon, F.*: Novum Organon. http://www.gmu.edu/departments/fld/CLASSICS/bacon.liber1.html, Abruf am 2005-01-19.

[Bage05] *Bager, J.*: Aufsteiger - Websites mit dem Internet Business Promoter suchmaschinengerecht aufbereiten. In: c't Magazin für Computertechnik, Bd.- H.9/2005, S.158-163.

[Bate84] *Bateson, G.*: Mind and Nature (Geist und Natur - eine notwendige Einheit). 3. Auflage Frankfurt am Main, 1984.

[BeKö02] *Becker, T.; König, E.*: Lexikonfreie Lemmatisierung für Substantive des Deutschen. In: Proceedings of Konvens 2002 - 6. Konferenz zur Verarbeitung natürlicher Sprache. http://konvens2002.dfki.de/cd/pdf/17V-Becker.pdf Saarbrücken, 2002, S.--.

[BeLi99] *Berry, M.; Linoff, G.*: Mastering Data Mining: The Art and Science of Customer Relationship Management. Hoboken, 1999.

[Berg+03] *Bergmann, R. u.a.*: Developing Industrial Case-Based Reasoning Applications - The INCREA Methodology. Lecture Notes in Artificial Intelligence Berlin, 2003.

[Bibel] *o.V.*: Das erste Buch Mose (Genesis) - Im fruchtbaren Garten. Basel und Gießen, 2003.

[Biem+03] *Biemann, C.; Bordag, S.; Quasthoff, U.*: Lernen paradigmatischer Relationen auf iterierten Kollokationen. In: LDV-Forum, Bd.½ H.19/2003, S.103-111.

[Boni90] *Bonitz, H.*: Aristoteles Metaphysik (erste Philosophie) - Übersetzung aus dem Griechischen. http://aristoteles.de.ki/, Abruf am 2004-10-18.

[Bose+92] *Boser, B.E. u.a.*: A Training Algorithm for Optimal Margin Classifiers. In: Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory. New York, 1992, S.144-152.

[BoTi+03] *Borgelt, C.; Timm, H.; Kruse, R.*: Unsicheres und vages Wissen. In: Handbuch der künstlichen Intelligenz. München, 2003, S.291-348.

[Bühl65] *Bühler, K.*: Sprachtheorie: Die Darstellungsform der Sprache. 2. Auflage Jena, 1965.

[Bußm02] *Bußmann, H.*: Lexikon der Sprachwissenschaft. Stuttgart, 2002.

[Cali98] *Califf, M. E.*: Relational Learning Techniques for Natural Language Information. http://www.cs.utexas.edu/users/ml/papers/rapier-dissertation-98.pdf, Abruf am 2005-07-06.

[CaSo03] *Cancho, R.F.; Sole, R.V.*: Least effort and the origins of scaling in human language. In: Proceedings of the National Academy of Sciences, Bd.3 H.10/2003, S.788-791.

[Chom02] *Chomsky, N.*: Syntaktic Structures. Studiengang Informationswirtschaft Berlin, 2002.

[Chom95] *Chomsky, N.*: Thesen zur Theorie der generativen Grammatik (Topics in the theory of generative grammar). 2. Auflage Weinheim, 1995.

[Chu+03] *Chung, W.; Chen, H.; Nunamaker, J.F.*: Business Intelligence Explorer: A Knowledge Map Framework for Discovering Business Intelligence on the Web. http://csdl.computer.org/comp/proceedings/hicss/2003/1874/01/187410010b.pdf, Abruf am 2005-02-01.

[CiHo+05] *Cimiano, P.; Hotho, A. ; Staab, S.*: Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. In: Journal of Artifical Intelligence Research, Bd.- H.24/2005, S.305-339.

[Cira+95] *Ciravegna, F.; Cedetta, N.*: Integrating Shallow and Linguistic Techniques for Information Extraction from Text. In: Topics in Artificial Intelligence. Lecture Notes in Artificial Intelligence, Bd.992, Heidelberg, 1995, S.127-138.

[CoNLL03] *o.V.*: Conference on Natural Language Learning - Language-Independent Named Entity Recognition (II). http://www.cnts.ua.ac.be/conll2003/ner/, Abruf am 2005-04-01.

[Coop91] *Cooper, W.S.*: Some inconsistencies and misnomers in probabilistic information retrieval. In: Proceedings of the 14th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval. SIGIR'91 Pittsburgh, 1991, S.57-61.

[CoSi99] *Collins, M.; Singer, Y.*: Unsupervised Models for Named Entity Classification. In: Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. College Park, 1999, S.100-110.

[Czap88] *Czap, H.*: Neue Ansätze in Terminologie und Wissenstechnik zur Unterstützung von Information und Kommunikation. In: Terminology and Knowledge Engineering. International Congress on Terminology and Knowledge Engineering Frankfurt am Main, 1988, S.212-223.

[Czap96] *Czap, H.*: Einführung in Wissensorganisation und Case Based Reasoning. In: Analogie in der Wissensrepresentation: Case-Based Reasoning und räumliche Modelle. Fortschritte in der Wissensorganisation, Bd.4, Frankfurt am Main, 1996, S.2-12.

[Dahl74] *Dahlberg, I.*: Grundlagen universaler Wissensordnung. Pullach, 1974.

[DaUt04] *Dauscher, P.; Uhtmann, T.*: Kurzeinfuhrung in die Informationstheorie. http://www.staff.uni-mainz.de/dauscher/al2/skripte/infotheo.pdf, Abruf am 2006-01-13.

[DCore06] *Dublin Core Metadata Initiative*: Dublin Core Metadata Element Set (ISO 15836-2003 siehe http://www.niso.org/international/SC4/n515.pdf). http://dublincore.org/documents/dces/, Abruf am 2006-05-02.

[Desc28] *Descartes, R.*: Regeln zur Leitung des Geistes (Regulae ad directionem ingenii). 2. Auflage Leipzig, 1920.

[Desc28a] *Descartes, R.*: Abhandlung über die Methode (Discours de la méthode). 2. Auflage Leipzig, 1920.

[Domi+01] *Dominique, J. u.a.*: Supporting Ontology Driven Document Enrichment within Communities of Practice. In: Proceedings 1st International Conference on Knowledge Capture. Elektronische Publikation Victoria B.C., 2001, S.--.

[Ebel+98] *Ebeling, W.; Freund, J.; Schweitzer, F.*: Komplexe Strukturen: Entropie und Information. Leipzig, 1998.

[Empo05] *Empolis GmbH*: Empolis orenge 4.0 Technology whitepaper. http://www.ovitas.com/PDF/orengeWhitepaper.pdf, Abruf am 2005-02-08.

[Ever+01] *Everitt, B.S.; Landau, S.; Leese, M.*: Cluster Analysis. London, 2001.

[FaNe98] *Faure, D.; Nedellec, C.*: Knowledge acquisition of predicate argument structures from technical texts using Machine Learning: the System ASIUM. In: Proceedings of the 11th European Workshop on Knowledge Acquisition, Modeling and Management. Lecture Notes In Computer Science, Bd.1621, New York, 1998, S.329-334.

[FaSt02] *Faatz, A.; Steinmetz, R.*: Ontology Enrichment with Texts from the WWW. In: Proceedings of 2nd Workshop on Semantic Web Mining. Helsinki, 2002, S.20-34.

[Ferb03] *Ferber, R.*: Information Retrieval - Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web. Heidelberg, 2003.

[Fisc89] *Fischer, G.*: Lineare Algebra. 9. Auflage Braunschweig, 1989.

[Fran02] *Frank, A.*: Einführung in die Computerlinguistik SS 2002 - Ambiguitäten, Kontext und corpusbasierte Sprachverarbeitung (I). http://www.ims.uni-stuttgart.de/Lehre/teaching/2002-SS/Einfuehrung-in-die-Computerlinguistik/ambiguity_context_corpora.ps, Abruf am 2004-03-03.

[Freg86] *Frege, F.L.G.*: Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens (gekürzter Nachdruck). 4. Auflage Berlin, 1986.

[FreMo72] *Freudenstein, R.; Moulton, W.G.*: Wie lernt man fremde Sprachen? (A linguistic guide to language learning). Dortmund, 1972.

[FrSc04] *Frank, U.; Schauer, H.*: Software für das Wissensmanagement: Einschlägige Systeme und deren Einführung. http://www.uni-koblenz.de/~iwi/publicfiles/PublikationenFrank/wisuWM.pdf, Abruf am 2004-10-10.

[Fugm92] *Fugmann, R.*: Theoretische Grundlagen der Indexierungspraxis. Gesellschaft für Wissensorganisation e.V. Frankfurt / Main, 1992.

[Gant04] *Ganter, B.*: Sprachvergleich - Qualitative Methoden - Vorlesungsskript SS 2004. http://www.math.tu-dresden.de/~ganter/glotto/QMBS04.pdf, Abruf am 2005-02-11.

[Gant05] *Ganter, B.*: Begriffe und Implikationen. www.math.tu-dresden.de/~ganter/psfiles/bwv.ps, Abruf am 2005-05-16.

[Gate05] *Cunningham, H.*: Information Extraction. http://gate.ac.uk/ie/, Abruf am 2005-04-25.

[GaWi99] *Ganter, B.; Wille, T.*: Formale Begriffsanalyse - Mathematische Grundlagen - in einer Aufbereitung von Wiebke Petersen. Heidelberg, 1999.

[GeCa03] *Geerts, G.L.; McCarthy, W.E.*: Tho Ontological Foundation of REA Enterprise Information Systems. http://www.msu.edu/user/mccarth4/Alabama.doc, Abruf am 2003-01-21.

[GerNe04] *GermaNet-Team*: GermaNet. http://www.sfs.uni-tuebingen.de/lsd/, Abruf am 2006-06-01.

[GoAm00] *Godoy, D.; Amandi, A.*: PersonalSearcher: An Intelligent Agent for Searching Web Pages. In: International Joint Conference, 7th Ibero-American Conference on AI, 15th Brazilian Symposium on AI, IBERAMIA-SBIA. Lecture Notes in Artificial Intelligence, Bd.1952, Heidelberg, 2000, S.43ff.

[Goet08] *von Goethe, J. W.*: Faust: Der Tragödie erster Teil. Leipzig, 1985.

[Gref92] *Grefenstette, G.*: Use of Syntactic Context to Produce Term Association Lists for Text Retrieval. In: Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, 1992, S.89-97.

[Gric77] *Grice, H.P.*: Bedeuten, meinen, intendieren. In: Paper. - Linguistic Agency, Univ. of Trier Trier, 1977, S.1-10.

[Grub93] *Gruber, T.R.*: A Translation Approach to Portable Ontology Specifications. In: Knowledge Acquisition, Bd.5 H.2/1993, S.199-221.

[Gurn97] *Gurney, K.*: An Introduction to Neural Networks. London, 1997.

[Hahn+01] *Hahn, U.; Honeck, M.; Schulz, S.*: A Search Engine for Morphologically Complex Languages. In: Advances in Intelligent Data Analysis. Lecture Notes in Computer Science, Bd.2189, London, 2001, S.73-83.

[Hand+03] *Handschuh, S.; Meyer, L. u.a.*: Ontomat-Annotizer. http://www.annotation.semanticweb.org/tools/ontomat, Abruf am 2003-05-11.

[Harm92] *Harman, D.*: Relevance Feedback Revisited. In: SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval. New York, 1992, S.1-10.

[HaSc98] *Hahn, U.; Schnattinger, K.*: Ontology Engineering via Text Understanding. In: Proceedings of the 15th IFIP World Computer Congress. Wien, 1998, S.429-442.

[Haus99] *Hausser, R.*: Foundations of Computational Linguistics - Man Machine Communication in Natural Language. Berlin, 1999.

[Hear92] *Hearst, M.A.*: Automatic Acquisition of Hyponyms from Large Text Corpora. In: Proceedings of COLING-92. Morristown, 1992, S.539-545.

[HeHo98] *Heckerman, D.; Horvitz, E.*: Inferring Informational Goals from Free-Text Queries: A Bayesian Approach. In: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence. San Fransisco, 1998, S.230-237.

[Herz06] *Herzog, G.*: Antw: WWW-Mail --> DIN 2342 / ISO 1087. Antwort vom 17.03.2006 auf persönliche Anfrage zur Definition "Begriff", Deutsches Institut für Normung, Berlin, 2006.

[Heye+01] *Heyer,G. u.a.*: Learning Relations using Collocations. In: Proc. IJCAI Workshop on Ontology Learning. Seattle, 2001, S.19-24.

[Hoff01] *Hoffmann, V.*: Grundüberlegungen zum Wissensmanagement. http://www.uni-hohenheim.de/i430a/lehre/veranst/download/skripten/pngk/pngk06.pdf, Abruf am 2005-11-01.

[HON05a] *o.V.*: Hintergrund: Was Adobe sich von der Macromedia-Übernahme verspricht. http://www.heise.de/newsticker/meldung/58709, Abruf am 2005-09-25.

[HON05b] *Bonnert, E.*: ISSCC: IBM und Sony präsentieren Cell-Prozessor. http://www.heise.de/newsticker/meldung/56139, Abruf am 2005-02-07.

[HoSt+03] *Hotho, A.; Staab, S.; Stumme, G.*: Ontologies Improve Text Document Clustering - The 2003 IEEE International Conference on Data Mining, Proceedings. http://www.aifb.uni-karlsruhe.de/WBS/aho/pub/hothoa_icdm_poster03.pdf, Abruf am 2005-02-16.

[HuCr78] *Hughes, G.E.; Cresswell, M. J.*: An introduction to modal logic. Berlin, 1978.

[INFO05] *INFODATA*: INFODATA Thesaurus. http://www.infodata-edepot.de/thesaurus/T_9.HTM#eMYg, Abruf am 2005-03-18.

[IsKa02] *Isozaki, H.; Kazawa, H.*: Efficient Support Vector Classifiers for Named Entity Recognition. http://acl.ldc.upenn.edu/C/C02/C02-1054.pdf, Abruf am 2005-10-21.

[KarTe01] *Karagiannis, D.; Telesko, R.*: Wissensmanagement - Konzepte der künstlichen Intelligenz und des Softcomputing. Lehrbücher Wirtschaftsinformatik, München, 2001.

[Klat02] *Klatt, S.*: Combining a rule-based tagger with a statistical tagger for annotating German texts. In: Proceedings of Konvens 2002 - 6. Konferenz zur Verarbeitung natürlicher Sprache. http://konvens2002.dfki.de/cd/pdf/15V-Klatt.pdf Saarbrücken, 2002, S.---.

[Klen05] *Klenner, M.*: Fomale Grundlagen der Linguistik - Vorlesungsskript SS 2005 Uni Zürich, Institut für Computerlinguistik. http://www.ifi.unizh.ch/cl/klenner/lehre/eclI-formal/formal.pdf, Abruf am 2005-08-04.

[Kluw92] *Kluwe, R.H.*: Gedächtnis und Wissen. In: Lehrbuch allgemeine Psychologie. , Bd.-, Bern, 1992, S.115-187.

[KöAlt86] *Köhler, R.; Altmann, G.*: Synergetische Aspekte der Linguistik. In: Zeitschrift für Sprachwissenschaft, Bd.- H.5/1986, S.253-265.

[Koho01] *Kohonen, T.*: Self Organizing Maps. 3. Auflage Berlin, 2001.

[Kolo93] *Kolodner, J.*: Case-Based-Reasoning. San Mateo, 1993.

[Kolz06] *Kolz,T.*: Methoden und Techniken des Dokumenten-Clustering; Diplomarbeit, Trier, 2006.

[KöRe98] *Königer, P; Reitmayer, W.*: Management unstrukturierter Informationen. Frankfurt, 1998.

[Kuhl77] *Kuhlen, R.*: Experimentelle Morphologie in der Informationswissenschaft. München, 1977.

[LäCl04] *Lämmel, U.; Cleve, J.*: Lehr- und Übungsbuch Künstliche Intelligenz. Leipzig, 2004.

[Lagu+05] *Lagus, K. u.a.*: WEBSOM map - comp.ai.neural-nets. http://websom.hut.fi/websom/comp.ai.neural-nets-new/html/root.html, Abruf am 2005-09-22.

[Lagu00] *Lagus, K.*: Text Mining with the WEBSOM; Dissertation, Espoo (Finnland), 2000.

[Lang04] *Lang, H.W.*: Vorlesung Compilerbau - Regulärer Ausdruck, reguläre Sprache. http://www.iti.fh-flensburg.de/lang/compbau/regular.htm, Abruf am 2005-01-10.

[Lav+04] *Lavelli, A. u.a.*: Distributional erm Representations: An Experimental Comparison. In: Proceedings of the thirteenth ACM international conference on Information and knowledge management. Washington DC, 2004, S.615-624.

[Lehm06] *Lehmann, C.*: Lexikalisierung und Grammatikalisierung. http://www.uni-erfurt.de/sprachwissenschaft/personal/lehmann/CL_Lehr/Morph&Syn/M&S_Lexikalisierung&Grammatikalisierung.html, Abruf am 2006-04-28.

[Lehn06] *Lehner, F.*: Wissensmanagement - Grundlagen, Methoden und technische Unterstützung. München, 2006.

[Leitn06] *Leitner, J.*: Extraktion von Ontologien aus natürlichsprachlichen Texten. Karlsruhe, 2006.

[Lepsk04] *Lepsky, K.*: Sprachengineering - Grundlagen und Methoden sprachverarbeitender Verfahren. http://www.iws.fh-koeln.de/institut/personen/lepsky/material/lehre/Skript-4-Sprachengineering.pdf, Abruf am 2005-01-31.

[Leve66] *Levenshtein, V.I.*: Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet Physics Doklady, Bd.10 H.8/1966, S.707-710.

[Lezi99] *Lezius, W.*: Morphy - Morphologie und Tagging für das Deutsche. http://www.lezius.de/wolfgang/morphy/, Abruf am 2005-01-31.

[Li+04] *Li, Y.; Bontcheva, K.; Cunningham, H.*: SVM Based Learning System For InformationExtraction. http://gate.ac.uk/sale/ml-ws04/mlw2004.pdf, Abruf am 2005-10-21.

[LoKa99] *Loucopoulus, P.; Kavakli, V.*: Enterprise Knowledge Management and Conceptual Modelling. In: Conceptual Modelling: Current Issues and Future Direction. Lecture Notes in Computer Science, Bd.1565, Heidelberg, 1999, S.123-143.

[Lore05] *Lorenz, S.*: Conceptual Patterns for Language Independent Information Extraction. In: Multikonferenz Wirtschaftsinformatik 2006. Track Wissensmanagement, Bd.2, Passau, 2005, S.447-460.

[Luhn58] *Luhn, H.P.*: Named Entity Recognition through Classifier Combination. In: Journal of Research and Development, Bd.2 H.2/1958, S.159-164.

[Lyons75] *Lyons, J.*: Einführung in die moderne Linguistik. München, 1975.

[Lyons77] *Lyons, J.*: Chomsky. 2. Auflage Hassocks, 1977.

[Lyons80] *Lyons, J.*: Semantik I. München, 1980.

[Maed+03] *Maedche, A.; Pekar, V.; Staab, S.*: Ontology Learning Part One - On Discovering Taxonomic Relations from the Web. In: Web Intelligence. Heidelberg, 2003, S.301-320.

[MaHü+03] *Mallot, H.A.; Hübner, W.; Stürzl, W.*: Neuronale Netze. In: Handbuch der künstlichen Intelligenz. München, 2003, S.73-124.

[Mande53] *Mandelbrot, B.B.*: An information theory of the statistical structure of language. In: Communication Theory. New York, 1953, S.503-512.

[MaRa05] *Manning, C.; Raghavan, P.*: Text Information Retrieval, Mining, and Exploitation - Web Search and Mining Winter 2005. http://www.stanford.edu/class/cs276b/handouts/lecture3.pdf, Abruf am 2005-06-25.

[MaSch00] *Manning, C. D.; Schütze, H.*: Foundations of statistical natural language processing. 2. Auflage Cambridge, 2000.

[Maur03] *Maurer, U.*: Information und Kommunikation - Vorlesungsskript der ETH Zürich. http://graphics.ethz.ch/teaching/infotheory/Downloads/skript.pdf, Abruf am 2005-06-24.

[Maur93] *Maurer, U.*: The Role of Information Theorie in Cyrptographie. In: Proc. of 4th IMA Conference on Cryptography and Coding. Southend-on-Sea, 1993, S.49-71.

[May+03] *Mayfield,J.; McNamee, P.; Piatko, C.*: Named Entity Recognition using Hundreds of Thousands of Features. In: - Proceedings of CoNLL-03. Edmonton, 2003, S.184-187.

[McCa05] *McCallum, A.*: Information Extraction: Distilling Structured Data from Unstructured Text. In: ACM Queue, Bd.3 H.9/2005, S.49-55.

[Mill+03] *Miller, G.A. u.a.*: WordNet - a lexical database for the English language. http://www.cogsci.princeton.edu/~wn/, Abruf am 2003-05-11.

[Mill65] *Mill, J. S.*: A System of Logic. London, 1865.

[Minsk75] *Minsky, M.*: A Framework for representing Knowledge. In: The Psychology of Computer Vision. New York, 1975, S.211-277.

[Mitch97] *Mitchell, T.M.*: Machine Learning. McGraw-Hill series in computer science New York, 1997.

[Mönc03] *Mönch, E.*: SemanticMiner - Ontology-based Knowledge Retrieval. In: Journal of Universal Computer Science, Bd.9 H.7/2003, S.682-692.

[Müll05] *Müller, S.*: Zur Nutzbarkeit konnektionistischer Methoden für die Klassifikation von Eigennamen; Diplomarbeit, Trier, 2005.

[Munr+03] *Munro, R.; Ler, D.; Patrick, J.*: Meta-Learning Orthographic and Contextual Models for Language Independent Named Entity Recognition. In: Proceedings of CoNLL-03. Edmonton, 2003, S.192-195.

[Net+00] *Neto, J. L. u.a.*: Generating Text Summaries through the Relative Importance of Topics. In: International Joint Conference, 7th Ibero-American Conference on AI, 15th Brazilian Symposium on AI, IBERAMIA-SBIA. Lecture Notes in Artificial Intelligence, Bd.1952, Heidelberg, 2000, S.300ff.

[Nohr00] *Nohr, H.*: Automatische Dokumentindexierung - eine Basistechnologie für das Wissensmanagement. In: Arbeitspapiere für das Wissensmanagement. Studiengang Informationswirtschaft, Bd.2/2000, Stuttgart, 2000, S.5f.

[NoTa95] *Nonaka, I.; Takeuchi, H.*: How Japanese Companies Create the Dynamics of Innovation. New York, 1995.

[Noy05] *Noy, N.*: Order from Chaos. In: ACM Queue, Bd.3 H.8/2005, S.42-49.

[Onto04] *Popov, B. u.a.*: KIM - Semantic Annotation Platform. http://www.ontotext.com, Abruf am 2005-03-05.

[Part+90] *Partee, B. H.; ter Meulen, A.; Wall, R.*: Mathematical Methods in Linguistics. Dordrecht, 1990.

[Pole78] *von Polenz, P.*: Geschichte der deutschen Sprache. Berlin, 1978.

[Pro2000] *o.V.*: The Protegé-Project. http://protege.stanford.edu/, Abruf am 2004-11-01.

[PuSt+03] *Puppe, F.; Stoyan,H.; Studer, R.*: Knowledge Engineering. In: Handbuch der künstlichen Intelligenz. München, 2003, S.599-641.

[Pyly73] *Pylyshyn, Z.W.*: What the mind's eye tells the mind's brain: A critique of mentalimagery. In: Psycological Bulletin, Bd. H.80/1973, S.1-24.

[Quas+02] *Quasthoff, U.; Biemann,C.; Wolff, C.*: Named Entity Learning and Verification: Expectation Maximization in Large Corpora. In: Proceedings of CoNLL-2002. New Brunswick, 2002, S.8-14.

[Quil85] *Quillian, M.R.*: Word Concepts: A Theory and Simulation of Some Basic Semantic Capabilities. In: Readings in Knowledge Representation. , Bd., Los Altos, 1985, S.97-118.

[QuWo02] *Quasthoff, U.; Wolff, C.*: Text-based Knowledge Acquisition for Ontology Engineering. In: Proceedings of Konvens 2002 - 6. Konferenz zur Verarbeitung natürlicher Sprache. Saarbrücken, 2002, S.147-154.

[Rabi89] *Rabiner, L.R.*: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In: Proceedings of the IEEE, Bd.77 H.2/1989, S.257-286.

[RaWa97] *Ravin, Y.; Wacholder, N.*: Extracting Names from Natural Language Text. http://domino.watson.ibm.com/library/cyberdig.nsf/a3807c5b4823c53f85256561006324be/0dd2ab07e511e076852565930072815d/$FILE/8648.ps.gz, Abruf am 2005-08-04.

[RDFS03] *o.V.*: Resource Description Framework (RDF). http://www.w3.org/RDF/, Abruf am 2003-04-01.

[ReiMa02] *Reimann-Rothmeier, G.; Mandl, H.*: Stichwort 'Wissen'. In: Lexikon der Psychologie. Band 5 Heidelberg, 2002, S.7-9.

[Reut00] *o.V.*: Reuters Corpus Volume 1, English LanguageFormat version 1, 1996-08-20 to 1997-08-19. http://trec.nist.gov/data/reuters/reuters.html, Abruf am 2005-08-15.

[Reut05] *o.V.*: Reuters Corpus Volume 2, Multilingual CorpusFormat version 1, 1996-08-20 to 1997-08-19. http://trec.nist.gov/data/reuters/reuters.html, Abruf am 2005-08-15.

[Röss04] *Rössler, M.*: Corpus Based Learning of Lexical Resources for German Named Entity Recognition - 4th International Conference on Language and Evaluation. http://cl.informatik.uni-duisburg.de/roessler/lrec2004.pdf, Abruf am 2005-04-24.

[RuNo04] *Russel, S.; Norvig,P.*: Künstliche Intelligenz Ein moderner Ansatz. 2. Auflage München, 2004.

[RuSm+87] *Rummelhart, D.E. u.a.*: Schemata and Sequential Thought Processes in PDP Models. In: Parallel distributed processing. Explorations in the microstructure of cognition. Psychological and Biological Models Cambridge, 1987, S.7-57.

[SaMcG83] *Salton, G.; McGill, M.J.*: Introduction to modern Information Retrieval. New York, 1983.

[Schm94] *Schmid, H.*: TreeTagger - ein sprachunabhängiger Wortart-Tagger. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger-de.html, Abruf am 2005-01-31.

[Schmi02] *Schmidt, S.*: Reguläre Ausdrücke. In: Perl für Profis. Bausteine fürs Web und fortgeschrittene Anwendungen Heidelberg, 2002, S.27-33.

[Schu95] *Schukat-Talamazzini, E.*: Automatische Spracherkennung Statistische Verfahren der Musteranalyse. Braunschweig, 1995.

[Schul81] *Schulz von Thun, F.*: Miteinander Reden 1 - Störungen und Klärungen. Reinbek, 1981.

[Shan48] *Shannon, C.E.*: A Mathematical Theory of Communication. In: The Bell System Technical Journal, Bd. H./1948, S.379-423.

[ShMa95] *Shardanand, U.; Maes, P.*: Social Information Filtering: Algorithms for Automating "Word of Mouth". In: Proceedings of the CHI '95. New York, 1995, S.210-217.

[Sing+99] *Singh, L. u.a.*: An Algorithm for Constrained Association Rule Mining in Semi-structured Data. In: Methodologies for Knowledge Discovery and Data Mining. Lecture Notes in Artificial Intelligence, Bd.1574, Berlin, Heidelberg, New York, 1999, S.148ff.

[SmMe81] *Smith, E.E.; Medin, D.L.*: Categories and Concepts. Cambridge, 1981.

[Sowa84] *Sowa, J.F.*: Conceptual Structures - Information Processing in Mind and Machine. Reading, 1984.

[Staab02] *Staab, S.*: Wissensmanagement mit Ontologien und Metadaten. In: Informatik Spektrum, Bd.25 H.3/2002, S.194-209.

[Stein+64] *Steiner, H.G. u.a.*: Das Fischer Lexikon Mathematik 1. Frankfurt am Main, 1964.

[StHa03] *Staab, S.; Handschuh, S.*: Semantic Annotation for the Semantic WebFrontiers in Artificial Intelligence and Applications. Amsterdam, 2003.

[Strö04] *Strötgen, R.*: ASEMOS. Weiterentwicklung der Behandlung semantischer Heterogenität. In: Informationen zwischen Kultur und Marktwirtschaft. Konstanz, 2004, S.269-281.

[Stum+01] *Stumme, G. u.a.*: Conceptual Clustering with Iceberg Concept Lattices - Universität Dortmund. http://www.aifb.uni-karlsruhe.de/WBS/Publ/2001/gst-fgml01.ps, Abruf am 2005-05-16.

[Titt03] *Tittmann, P.*: Graphentheorie - Eine anwendungsorientierte Einführung. Leipzig, 2003.

[Titz93] *Titzmann, M.*: Strukturale Textanalyse - Theorie und Praxis der Interpretation. 3. Auflage München, 1993.

[TREC04] *o.V.*: Question Answering Collections. http://trec.nist.gov/data/qa.html, Abruf am 2004-11-25.

[Trier73] *v. Trier, J.*: Aufsätze und Vorträge zur Wortfeldtheorie. Den Haag, 1973.

[Trier73a] *v. Trier, J.*: Der deutsche Wortschatz im Sinnbezirk des Verstandes. Heidelberg, 1973.

[TsiLa02] *Tsikrika, T.; Lalmas, M.*: Combining Web Document Representations in a Bayesian Inference Network Model Using Link and Content-Based Evidence. In: Advances in Information Retrieval. 2291 -24th BCS IRSG European Colloquium on IR Research Heidelberg, 2002, S.53-72.

[Tuomi00] *Tuomi, I.*: Data is more than Knowledge: Implications of the Reversed Hierarchy for Knowledge Management and Organizational Memory. In: Journal of Management Information Systems, Bd.16 H.3/2000, S.103-118.

[UnSh05] *University of Sheffield*: GATE - General Architecture for Text Engineering - fact file. http://www.aktors.org/technologies/gate/, Abruf am 2005-03-01.

[Usch+98] *Uschold, M. u.a.*: The Enterprise Ontology. http://www.aiai.ed.ac.uk/publications/documents/1998/98-ker-ent-ontology.ps, Abruf am 2003-01-21.

[Vapn95] *Vapnik, V.N.*: The Nature of Statistical Learning Theory. 2. Auflage Heidelberg, 1995.

[Var+02] *Vargas-Vera, M. u.a.*: MNM: Ontology-Driven Tool for Semantic Markup. In: The 13th International Conference on Knowledge Engineering and Management (EKAW 2002). Lecture Notes in Artificial Intelligence, Bd.2473, Heidelberg, 2002, S.379-391.

[Volk79] *Volkmann-Schluck, K.-H.*: Die Metaphysik des Aristoteles. Frankfurt am Main, 1979.

[W3C01] *WWW-Consortium*: W3C Semantic Web. http://www.w3.org/2001/sw/, Abruf am 2006-06-13.

[W3C04] *WWW-Consortium*: OWL Web Ontology Language Overview. http://www.w3.org/TR/2004/REC-owl-features-20040210/, Abruf am 2006-06-13.

[Wede92] *Wedekind, H.*: Objektorientierte SchemaentwicklungEin kategorialer Ansatz für Datenbanken und Programmierung. Reihe Informatik Mannheim, 1992.

[Whew58] *Whewell, W.*: History of Scientific Ideas. London, 1858.

[WiHo60] *Widrow, B.; Hoff, M. E.*: Adaptive switching Circuits. In: 1960 IRE WESCON Convention Record. New York, 1960, S.96-104.

[Wil63] *Wilson, J.*: Thinking with Concepts - Begriffsanalyse. Cambridge / Stuttgart, 1963.

[Witt18] *Wittgenstein, L.*: Tractatus logico-philosophicus - Deutsche Ludwig Wittgenstein Gesellschaft. http://www.kfs.org/~jonathan/witt/tde.html, Abruf am 2004-10-15.

[Witt53] *Wittgenstein, L.*: Philosophical Investigations. New York, 1953.

[Wolf01] *Wolfertz, K.*: Wissensmanagement bei Beratern mit Fuzzy Systems. In: Wirtschaftsinformatik, Bd.43 H.5/2001, S.457-466.

[WrMo+03] *Wrobel, S.; Morik, K.; Joachims, T.*: Maschinelles Lernen und Data Mining. In: Handbuch der künstlichen Intelligenz. München, 2003, S.517-598.

[Zaun99] *Zaun, D. P.*: Künstliche neuronale Netze und Computerlinguistik - Linguistische Arbeiten 406. Tübingen, 1999.

[Zipf32] *Zipf, G.K.*: Selected studies of the principle of relative frequency in language. Cambridge, 1932.