# Definitionsextraktion aus Urteilstexten

**Dissertation zur Erlangung des akademischen Grades eines Doktors der Philosophie der Philosophischen Fakultäten der Universität des Saarlandes**

vorgelegt von
Stephan Walter
aus Hannover

Tag der letzten Prüfungsleistung: 1. Dezember 2010

Der Dekan: Prof. Dr. Erich Steiner

Berichterstatter: Prof. Dr. Manfred Pinkal
Prof. Dr. Maximilian Herberger

Meinem Vater

# Danksagung

Zuallererst möchte ich mich bei meinem Doktorvater Manfred Pinkal bedanken, ohne dessen Unterstützung, Rat, Vertrauen und nicht zuletzt Geduld diese Arbeit nie zu Stande gekommen wäre. Er hat mir große Freiheit gelassen, meine Ideen auch dann zu verfolgen, wenn nicht direkt ersichtlich war, zu welchem Ergebnis sie führen. Dabei hat er aber auch dafür gesorgt, dass sie sich zuerst zu einem Thema und schließlich auch zu einem Text verfestigen. Sein Lehrstuhl bietet ein großartiges Umfeld für selbstbestimmtes und trotzdem eingebundenes wissenschaftliches Arbeiten.[1]

Bei Maximilian Herberger bedanke ich mich für viele wichtige Beiträge und Hinweise zu den rechtswissenschaftlichen Aspekten meiner Arbeit, besonders aber für das große allgemeine Interesse, das er der ganzen thematischen Breite und dem interdisziplinären Ansatz des Forschungsvorhabens entgegengebracht hat.

In Saarbrücken und anderswo habe ich vielen Leuten dafür zu danken, dass sie mir mit Kommentaren, Hinweisen, in Gesprächen und Diskussionen geholfen haben, meine Gedanken zu entwickeln und diese Arbeit zu realisieren: Peter Adolphs, Christian Braun, Aljoscha Burchardt, Ralph Dornis, Gerd Fliedner, Anette Frank, Elena Karagjosova, Martin Kerz, Alexander Koller, Martin Küchler, Stella Neumann, Marco Pennacchiotti, Sebastian Padó, Werner Saurer, Tatjana Scheffler, Torgrim Solstad, Stefan Thater, Dimitra Tsovaltzi, Magdalena Wolska. Klaas Schmidt und Michael Wiegand haben mir als studentische Hilfskräfte nicht nur viel Arbeit abgenommen, sondern auch oft mit interessanten Anmerkungen und Hinweisen weitergeholfen.

Monika Rathert hat meine Arbeit in der wichtigen und schwierigen Phase der ersten schriftlichen Entwürfe mit großem Interesse und vielen wertvollen Kommentaren begleitet.

Sehr ermutigend war für mich auch das Interesse, das meinen Ideen verschiedentlich “von praktischer Seite” entgegengebracht wurde. In diesem Zusammenhang möchte ich mich vor allem bei Franz Kummer und Mario Krauss bedanken.

[1] Ein Großteil der in dieser Arbeit dokumentierten Forschungen fand im Rahmen des von der Deutschen Forschungsgemeinschaft geförderten Projektes CORTE statt (*Computerlinguistische Methoden für die Rechtsterminologie*, Geschäftszeichen PI 154/10-1)

Für weit über Administratives hinausgehende praktische Hilfe bedanke ich mich bei Christoph Clodo, Ursula Kröner, Angelika Maurus, Bobbye Pernice und Helga Riedel.

Meiner Familie danke ich dafür, dass sie mich von Anfang an und auch in einer sehr schweren Zeit in jeder Form unterstützt und mir immer Verständnis entgegengebracht hat.

Mein letzter, aber um nichts geringerer Dank gebührt Hanyi, die wohl alle Aspekte meines Arbeitens besonders deutlich zu spüren bekommen hat. Sie hat in jeder Phase, auch als kein Ende abzusehen schien, vollkommen hinter mir gestanden.

# Inhaltsverzeichnis

# Abbildungsverzeichnis

# Tabellenverzeichnis

# Einleitung

Rechtlich relevante Terminologie ist nur zum Teil durch Definitionen in Gesetzestexten festgelegt. Zusätzlich findet im Rahmen der Rechtsprechung eine permanente lokale, fallbezogene Präzisierung und Konkretisierung der verwendeten Begriffe statt. Gerichtsentscheidungen stellen deshalb neben Gesetzestexten eine zweite wichtige Quelle konzeptuellen Wissens für die Rechtsprechungspraxis dar. Um dieses Wissen automatisch erschließen zu können, müssen Definitionen in Entscheidungstexten mit hoher Präzision und Abdeckung aufgefunden werden. Diese Arbeit befasst sich mit der automatischen Extraktion und Analyse von Definitionen in deutschsprachigen Gerichtsurteilen.

Methoden und Techniken für solche Zwecke sind in jüngerer Zeit unter dem Begriff *Text Mining* aktuell geworden. Aufgrund der Besonderheiten der juristischen Domäne unterscheidet sich die Zielsetzung der Arbeit jedoch in wichtigen Punkten von typischen Fragestellungen in diesem Bereich.

Zunächst einmal können bei der Definitionssuche in Entscheidungstexten nicht in gleichem Maße Redundanzen genutzt werden wie beispielsweise bei der Suche nach Konzepten oder Relationen im WWW. Dort kommen Informationen in aller Regel an einer Vielzahl von Stellen in unterschiedlichsten linguistischen Realisierungen vor. Es bestehen daher gute Voraussetzungen für den Einsatz statistischer Methoden. Dagegen ist im juristischen Kontext für einen konkreten Rechtsfall unter Umständen eine Definition relevant, die nur in einem einzigen Urteilstext enthalten ist. Zusätzlich kann es auf spezifische Details der Formulierung ankommen. Es muss daher oft treffsicher eine eng eingegrenzte Passage in einem einzelnen Dokument aufgefunden werden. Hilfreich kann hierbei allerdings die ausgeprägte Makrostruktur von Entscheidungstexten sein, die im Gegensatz zu vielen anderen Textsorten strikten Konventionen folgt und in hohem Maße explizit gemacht wird.

Gleichzeitig enthalten Definitionen desselben Begriffs in verschieden Urteilen oft einander ergänzende und aufeinander bezogene Informationen (z.B. Präzisierungen eines übergreifenden Konzepts für unterschiedliche Fallklassen, etwa des Werkbegriffs im Urheberrecht für den Bereich der Literatur, der Architektur und der Gebrauchskunst). Neben Zielgenauigkeit ist deshalb bei der Definitionsextraktion auch eine hohe Abdeckung wichtig, um die konzeptuelle Gesamtstruktur von Rechtsbegriffen möglichst detailliert erschließen zu können.

Unter solchen Rahmenbedingungen stoßen statistische Verfahren an die Grenzen ihrer Leistungsfähigkeit. Sie können jedoch zum einen durch Expertenwissen ergänzt werden. Zum anderen ist es in dieser Situation sinnvoll, verhältnismäßig tiefe linguistische Information in den Suchprozess einzubeziehen, selbst wenn diese nur mit recht hohem Verarbeitungsaufwand ermittelt werden kann. Im Fall von Rechtstexten werden dafür computerlinguistische Analysewerkzeuge benötigt, die trotz der Eigenheiten der Rechtssprache (z.B. ihrer syntaktischen Komplexität und der verwendeten Fachterminologie) robust Ergebnisse liefern.

Hauptbeitrag meiner Arbeit ist ein linguistisch informiertes Definitionsextraktionsverfahren, das Expertenwissen in Form von Suchmustern nutzt, die auf der Grundlage einer systematischen Analyse eines Rechtsprechungskorpus spezifiziert wurden. Es stützt sich auf die Analysen einer spezialisierten sprachtechnologischen Verarbeitungskette für Rechtstexte, die ich auf der Basis des in Fliedner (2007) genutzten PreDS-Parsers entwickelt habe. Durch Hinzunahme datengetrieben arbeitender Optimierungsschritte verbessere und erweitere ich die Definitionssuche bis zur praktischen Anwendungsreife.

Diese Forschungen konnten natürlich nicht durchgeführt werden, ohne grundlegende rechtswissenschaftliche Zusammenhänge mit zu berücksichtigen. Ich räume deshalb auch der Erläuterung dieser konzeptuellen Grundlagen im Folgenden großzügig Platz ein (besonders im ersten und zweiten Kapitel). Das Resultat bleibt selbstverständlich eine computerlinguistische Arbeit und vor allem, trotz kompetenter und engagierter Unterstützung durch Rechtswissenschaftler, das Produkt eines juristischen Laien. Ich würde mich jedoch freuen, wenn die Arbeit durch die thematische Erweiterung auch für die (computer)linguistische Leserschaft einen interessanten Aspekt hinzugewonnen hätte.

Noch eine weitere Anmerkung möchte ich gerne vorab zur Einordnung meiner Resultate machen: Thematisch und methodologisch betritt die Arbeit an mehreren Stellen Neuland. Ich bin mir daher bewusst, nicht überall umfassende und abschließende Ergebnisse präsentieren zu können. Allerdings hoffe ich dort, wo Fragen offen bleiben, zumindest Ansatzpunkte für weitere interdisziplinäre Forschungen aufgezeigt zu haben.

## Aufbau der Arbeit

Die Arbeit gliedert sich in zwei Teile. Der erste Teil ist einer Untersuchung des Phänomens *Definitionen in der Rechtssprache* gewidmet. Er umfasst die ersten beiden Kapitel. Der zweite Teil befasst sich – auf der Basis der Ergebnisse der ersten beiden Kapitel – mit der Konzeption, Umsetzung und Optimierung eines

Verfahrens zur *automatischen Extraktion von Definitionen* aus deutschsprachigen Urteilstexten.

**Kapitel 1** enthält – zur Überprüfung durch den juristischen und als Hintergrund für den computerlinguistischen Leser – die Darstellung meines Verständnisses des juristischen Auslegungsvorgangs. Ich erläutere, wie Entscheidungsbegründungen an wesentlichen Stellen auf die semantische Ausdifferenzierung zentraler Konzepte durch definitorische Elemente gestützt werden. Dabei begründe ich die generelle Relevanz der in dieser Arbeit untersuchten Thematik und arbeite konzeptuelle Unterschiede zu ähnlichen Fragestellungen in anderen Anwendungsdomänen heraus.

**Kapitel 2** charakterisiert zunächst einen weit gefassten Definitionsbegriff, der sich für den juristischen Zusammenhang als sinnvoll erweist. Er umfasst neben vollständigen Definitionen verschiedene Typen unvollständiger Begriffsbestimmungen. Dabei bezieht er insbesondere die Aspekte *textuelle Einbindung* und *argumentative Funktion* solcher Aussagen mit ein. Es schließt sich eine Untersuchung und Systematisierung der Bandbreite der Formulierungsmuster an, die in der Rechtssprache für Definitionen verwendet werden. Den Abschluss des Kapitels bildet die Auswertung einer Annotationsstudie. Sie bietet Anhaltspunkte zum Übereinstimmungsgrad menschlicher Leser bei der Identifikation rechtssprachlicher Definitionen und liefert zudem einen Goldstandard für die Evaluation des Verfahrens zur automatischen Definitionsextraktion, das im zweiten Teil der Arbeit vorgestellt wird.

**Kapitel 3** stellt den Stand der informationstechnologischen Forschungen dar, die für das Thema der Arbeit relevant sind. Nach einem knappen Überblick über Ansätze zum textbasierten automatischen Informationszugriff im allgemeinen stehen Aufgabenstellungen im Vordergrund, bei denen die automatische Identifikation und / oder Verarbeitung von Definitionen eine besondere Rolle spielt. Das Kapitel schließt mit einem Überblick zu juristischen Anwendungen moderner Informationstechnologie und diskutiert kurz die Gründe, aus denen in dieser Domäne Text Mining und ähnliche Techniken noch von vergleichsweise geringer Bedeutung sind.

**Kapitel 4** wendet sich dem technischen Hauptthema der Arbeit zu: der Entwicklung eines Verfahrens zur automatischen Identifikation von Definitionen in Urteilsbegründungen und seiner Umsetzung im CORTE-System.[2] Der Schwer-

[2] benannt nach dem Projekt CORTE (*Computerlinguistische Methoden für die Rechtsterminologie*), in dessen Rahmen ein großer Teil der hier beschriebenen Arbeiten stattfand.

punkt des Kapitels liegt auf der Realisierung einer Vorverarbeitungskette für diesen Zweck mit alternativen Verarbeitungswegen auf der Basis flacher und tiefer Analysekomponenten. Den zweiten Teil des Kapitels bildet die Beschreibung der Datengrundlage, die für die Umsetzung und Erprobung des CORTE-Systems zum Einsatz kommt. Ein Exkurs behandelt verschiedene linguistische Charakteristika der Rechtssprache. Sie konnten auf dieser Datengrundlage (zum Teil erstmalig) umfassend empirisch belegt werden. Besondere Beachtung kommt dabei Phänomenen mit Auswirkungen auf die sprachtechnologische Verarbeitung von Rechtstexten zu.

**Kapitel 5** befasst sich mit dem eigentlichen Definitionsextraktionsverfahren. Dieses identifiziert und analysiert regelbasiert Definitionen in Entscheidungstexten, die die im vorigen Kapitel erläuterte Vorverarbeitungskette durchlaufen haben. Es nutzt Suchmuster, die direkt zu den in Kapitel 2 identifizierten definitionstypischen Formulierungen korrespondieren. Die Performanz der Definitionsextraktion und -analyse mit diesen Mustern wird anhand des (ebenfalls in Kapitel 2 erzeugten) Goldstandards und weiterer Referenzressourcen ausgewertet. Dabei zeigt sich deutlich der positive Effekt der Nutzung tiefer linguistischer Information im Suchprozess. Auf der Grundlage einer Stichprobe der aufgetretenen Fehler werden dann Ansatzpunkte für die Optimierung des Extraktionsverfahrens ermittelt.

**Kapitel 6** verfolgt verschiedene Ansätze zur Verbesserung der Ergebnisqualität der Definitionsextraktion. Naheliegende, jedoch wissensintensive und eher punktuelle Maßnahmen, wie die Verwendung von Filtern und die manuelle Spezifikation zusätzlicher Suchmuster, erweisen sich als nicht effektiv. Für die praktische Verwendung sind sie zudem zu arbeitsaufwändig. Zu deutlichen Verbesserungen führen dagegen datengestützte Optimierungsansätze. Den stärksten Effekt erzielt ein Rankingverfahren auf der Basis heuristischer Definitionsindikatoren in Kombination mit einem Bootstrapping-Ansatz zum automatischen Erwerb neuer Suchmuster.

**Kapitel 7** beschließt die Arbeit mit einer Zusammenfassung der erzielten Ergebnisse sowie dem Ausblick auf zwei Anwendungsperspektiven: die direkte Nutzung des Extraktionssystems zum definitionsbasierten Zugriff auf eine Entscheidungssammlung und die Verwendung von Extraktionsergebnissen bei der induktiven Modellierung juristischen Wissens.

# Kapitel 1

# Definitionen in der Rechtssprache

Die Rechtsprechung kann nur dann rechtsstaatlichen Ansprüchen genügen, wenn sichergestellt ist, dass die Normen des Rechtssystems einheitlich und nachvollziehbar angewandt werden. Gesetzgebung und Rechtsprechung finden in natürlicher Sprache statt. In der Rechtsanwendung müssen Normen deshalb *gemäß ihrer sprachlichen Bedeutung* auf die gesellschaftliche Realität bezogen werden. Eine einheitliche und nachvollziehbare Rechtsanwendung setzt somit einen einheitlichen und nachvollziehbaren Sprachgebrauch (zumindest hinsichtlich der rechtlich relevanten Begriffe) voraus.

Aus diesem Grund werden in vielen Gesetzestexten zentrale Begriffe durch Definitionen festgelegt. Dadurch wird der juristische Sprachgebrauch jedoch nur zum Teil determiniert. Um die global fixierten Rechtsvorschriften auf eine nicht im Vorhinein überschaubare Vielfalt von Fällen anwenden zu können, ist zusätzlich eine permanente lokale, fallbezogene Präzisierung und Konkretisierung der rechtssprachlichen Begrifflichkeiten erforderlich. Diese wird methodengeleitet im Rahmen der *Auslegung* vorgenommen und in den *Entscheidungsgründen* zu jedem Fall schriftlich dokumentiert. Nicht nur in Gesetzestexten, sondern auch in Gerichtsentscheidungen sind deshalb begriffsbestimmende Äußerungen wesentliche Bestandteile. Auch diese Definitionen bleiben jedoch oft über den Einzelfall hinaus relevant. Urteilsbegründungen stellen somit neben Gesetzestexten eine zweite wichtige Quelle begrifflichen Wissens für die Rechtsprechungspraxis dar.

In diesem Kapitel erläutern wir diese Zusammenhänge näher. Wir stellen zuerst in Grundzügen die Rolle der Auslegung im Gesamtprozess der Rechtsanwendung dar (1.1). Dann gehen wir auf die Gemeinsamkeiten und Unterschiede zwischen dem definitionsbasierten Terminologieaufbau in wissenschaftlich-technischen Fachsprachen und der Bestimmung von Rechtsbegriffen ein (1.2). Schließlich betrachten wir anhand von Textbeispielen im Detail, wie in Urteilsbegründungen mit Begriffsbestimmungen argumentiert und dabei die Semantik für die Argumentation zentraler Begriffe im Textzusammenhang ausdifferenziert wird (1.3).

## 1.1 Normen und ihre Auslegung

Recht dient der Ordnung und Regelung menschlichen Handelns. Es bestimmt durch ein Gefüge allgemeinverbindlicher Normen, wie die Mitglieder einer Gemeinschaft sich in den verschiedensten Situationen verhalten sollen. Konfliktsituationen zwischen Mitgliedern der Gemeinschaft werden anhand der Normen der Rechtsordnung entschieden, und Entscheidungen in solchen Situationen werden aufgrund dieser Normen durch rationale Argumentation begründet.[1]

Die Normen der Rechtsordnung sind sprachlich gefasst und – zumindest in vielen modernen Staatswesen – in Gesetzestexten schriftlich niedergelegt.[2] Auch die Begründung von Entscheidungen erfolgt in Form sprachlich gefasster Argumente, verschriftlicht in Urteilstexten. Ausgangspunkt sind dabei Rechtssätze, die einzelne normative Inhalte zum Ausdruck bringen und im einfachsten Fall direkt einem Gesetzestext entnommen werden können.

### 1.1.1 Logische Struktur von Normen

Der logischen Grundstruktur nach handelt es sich bei den durch einzelne Rechtssätze zum Ausdruck gebrachten normativen Inhalten um allquantifizierte Konditionale mit einem deontisch "gesollten" Konsequens. Das Antezedens beschreibt die sanktionierten Tatsachen (den *Tatbestand*), während das Konsequens die *Rechtsfolge* festlegt, die bei Vorliegen des Tatbestands eintreten soll. Mit den Symbolen einer deontisch erweiterten Prädikatenlogik lässt sich dies so darstellen:

$$\forall x(T(x) \implies \Box R(x))$$

Dieses Schema stellt eine starke strukturelle Vereinfachung der realen Verhältnisse dar.[3] In aller Regel ist der Tatbestand (und unter Umständen auch die

[1] Die *Gesetzesbindung* der Rechtsprechung ist in der deutschen Rechtsordnung gleich an mehreren Stellen in der Verfassung verankert, nämlich in Art. 20 Abs. 3 GG, Art. 97 Abs. 1 GG und – für das Strafrecht – noch in Art. 103 Abs. 2 GG. Die verschiedenen Prozessordnungen verpflichten Gerichte zudem zur expliziten Begründung ihrer Entscheidungen (z.B. § 313 ZPO, § 267 StPO)

[2] Die in dieser Arbeit gemachten Aussagen sind zunächst grundsätzlich nur auf den deutschen Rechtskreis mit seinem kodifizierten Rechtssystem zu beziehen. Was davon auf das z.B. im angelsächsischen Rechtskreis gängige *case law* übertragbar ist, kann nicht ohne eingehendere Untersuchungen festgestellt werden.

[3] Auch die logischen Grundlagen der Formalisierung von Rechtsnormen sind nicht unumstritten. Zunächst einmal kann grundsätzlich hinterfragt werden, ob Normen wahrheitswertfähig sind. Zumindest die Verwendung eines wahrheitsfunktionalen Konditionals in der Formalisierung wird oft als problematisch betrachtet, da es intuitiv unplausibel erscheint, Normen mit unerfüllbarem Tatbestand als logisch wahr zu betrachten (vgl. Weinberger (1979)). Andererseits erscheint die Annahme von Folgerungsbeziehungen wie den im Folgenden zu Grunde gelegten intuitiv auch für normative Aussagen plausibel ("Jørgensens Dilemma", vgl. Jørgensen (1937)).

Rechtsfolge) eine Verknüpfung mehrstelliger Prädikate, und die Allquantifikation erfolgt über mehrere Argumente. So legt beispielsweise § 524 Abs. 1 BGB fest:

> Verschweigt der Schenker arglistig einen Fehler der verschenkten Sache, so ist er verpflichtet, dem Beschenkten den daraus entstehenden Schaden zu ersetzen.

Der Tatbestand besteht hier aus einer konjunktiven Verknüpfung mehrerer *Tatbestandsmerkmale*:

1. Es hat eine Schenkung einer Sache durch einen Schenker an einen Beschenkten stattgefunden.
2. Die verschenkte Sache ist mit einem Fehler behaftet.
3. Der Schenker hat diesen Fehler verschwiegen.
4. Er hat dies arglistig getan.
5. Dem Beschenkten ist durch den Fehler ein Schaden entstanden.

Der zitierte § 524 Abs. 1 BGB zeigt zudem, dass Tatbestandsmerkmale sprachlich auch als Präsuppositionen, d.h. implizit durch den Gebrauch des bestimmten Artikels, eingeführt werden. Als solche können sie eventuell der Beschreibung der Rechtsfolge zu entnehmen oder indirekt zu erschließen sein. Präsupponiert sind hier Merkmal 1 (zu erschließen aus der definiten Verwendung der Ausdrücke *Schenker*, *Beschenkter* und *verschenkte Sache*) sowie Merkmal 5 (durch den Ausdruck *den . . . Schaden* in der Beschreibung der Rechtsfolge).

Die Rechtsanwendung erfordert also zunächst einmal die Auflösung der sprachlichen Komplexität der Normformulierungen. Ansonsten aber lässt das bisher Gesagte vermuten, dass konkrete Fälle vor allem durch die Anwendung logischer Schlussregeln gelöst werden: Durch Anwendung der Allbeseitigungsregel werden die allquantifizierten Variablen in Tatbestand und Rechtsfolge entsprechend dem zu entscheidenden Sachverhalt instantiiert. Aus § 524 Abs. 1 BGB wird so zum Beispiel eine partikuläre Aussage gewonnen, die sich auf eine konkrete Schenkung bezieht und auf einen Schaden referiert, der dem konkreten Beschenkten in dem zu beurteilenden Fall durch einen bestimmten Fehler der verschenkten Sache tatsächlich entstanden ist. Mittels des Modus ponens kann dann aus dieser partikulären Norm und der Beschreibung des Sachverhalts

---

Auf die Grundlagendiskussion zur Semantik normativer Aussagen kann im Rahmen dieser Arbeit nicht näher eingegangen werden.

auf die instantiierte Rechtsfolge geschlossen werden (im Beispiel also auf die Schadenersatzpflicht des konkreten Schenkers).

Praktisch können Fälle aber so gut wie nie auf so einfache und direkte Art anhand einzelner Normsätze gelöst werden. Dies liegt zu einem daran, dass für die meisten Fälle die anzuwendende Norm nicht in Form eines einzelnen vollständigen Rechtssatzes einem Gesetzestext entnommen werden kann. Sie muss erst unter Berücksichtigung einer (unter Umständen beträchtlichen) Anzahl verschiedener Textstellen gewonnen werden. Zum anderen können nur selten alle notwendigen Tatbestandsmerkmale in instantiierter Form direkt und eindeutig mit Elementen der Beschreibung des zu beurteilenden Sachverhalts zur Deckung gebracht werden. Dieser Bezug muss erst durch *Auslegung* des Tatbestands hergestellt werden.

### 1.1.2 Komplexität des Normengefüges

Nur in den seltensten Fällen findet sich in einem Gesetzestext genau ein einzelner Rechtssatz, der den zu entscheidenden Fall vollständig erfasst. Um im Einzelfall eine Norm zu erhalten, auf die sich eine Entscheidung stützen kann, müssen in der Regel Rechtssätze von verschiedenen Stellen eines Gesetzestexts, unter Umständen auch aus verschiedenen Gesetzestexten, zusammengetragen und in der richtigen Weise aufeinander bezogen werden.

#### (a) Vollständige und unvollständige Rechtssätze

Rechtssätze lassen sich nach ihrer Funktion bei der Lösung von Fällen einteilen in *vollständige* und *unvollständige Rechtssätze*. Vollständige Rechtssätze (wie der oben zitierte § 524 Abs. 1 BGB) enthalten sogenannte *Antwortnormen*, d.h. sie sind geeignet, in bestimmten Konfliktfällen selbständig die konkret gesuchte Rechtsfolge zu bestimmen und zu begründen. Unvollständige Rechtssätze sind entweder überhaupt nur im Zusammenhang mit vollständigen Rechtssätzen anwendbar, oder sie spezifizieren zumindest Rechtsfolgen, die sich nur im Zusammenhang mit anderen Normen auf den Fall selber auswirken.

Unvollständige Rechtssätze können in ganz unterschiedlicher Art auf vollständige Rechtssätze bezogen sein.[4] *Verweisungen* übernehmen beispielsweise Tatbestand oder Rechtsfolge anderer Rechtssätze (*Rechtsgrund-* bzw. *Rechtsfolgeverweis*). Andere unvollständige Rechtssätze ergänzen oder modifizieren den Inhalt vollständiger Rechtssätze, zum Beispiel in *erläuternder* Funktion (sog. *Hilfsnormen*) oder indem sie *Ausnahmen* regeln (sog. *Gegennormen*). Die im Folgenden wiedergegebenen Teile von §§ 528 und 529 BGB (Regelungen

[4] Die hier diskutierte Klassifikation orientiert sich an Schwacke (2003), 25 ff.; vgl. aber auch Larenz (1991), 257–264

zum *Rückforderungsanspruch bei Verarmung des Schenkers*) enthalten mehrere unvollständige Rechtssätze:

> § 528 BGB
>
> (1) Soweit der Schenker nach der Vollziehung der Schenkung außerstande ist, seinen angemessenen Unterhalt zu bestreiten (...) kann er von dem Beschenkten die Herausgabe des Geschenkes nach den Vorschriften über die Herausgabe einer ungerechtfertigten Bereicherung fordern. (...)
>
> (2) Unter mehreren Beschenkten haftet der früher Beschenkte nur insoweit, als der später Beschenkte nicht verpflichtet ist.
>
> § 529 BGB
>
> (1) Der Anspruch auf Herausgabe des Geschenkes ist ausgeschlossen, wenn der Schenker seine Bedürftigkeit vorsätzlich (...) herbeigeführt hat (...).

§ 528 Abs. 1 Satz 1 ist eine Verweisung auf die *Vorschriften über die Herausgabe einer ungerechtfertigten Bereicherung*. § 528 Abs. 2 enthält eine *erläuternde Hilfsnorm*. Sowohl der Tatbestand (mehrere Beschenkte) als auch die Rechtsfolge (beschränkte Haftung des später Beschenkten) sind vor dem Hintergrund der (entsprechend der Verweisung vervollständigten) Norm in § 528 Abs. 1 Satz 1 zu verstehen und nur im Zusammenhang mit dieser relevant. § 529 Abs. 1 stellt eine *rechtshindernde* Gegennorm zu § 528 Abs. 1 dar (Gegennormen können auch *rechtsvernichtend* sein, wenn sie in ein bereits bestehendes Recht eingreifen, oder sie können die *Durchsetzbarkeit* einer Rechtsfolge hindern). Es werden Ausnahmebedingungen angegeben, deren Rechtsfolge im Ausbleiben des Rückforderungsanspruchs bei Vorliegen der Tatbestandsvoraussetzungen von § 528 Abs. 1 besteht. Auch hier sind die Auswirkungen auf den konkreten Fall also indirekt, vermittelt durch die Antwortnorm, die sich gemäß der Verweisung in § 528 Abs. 1 ergibt.

#### (b) Zusammenwirken und Konkurrenz von Normen

Auch vollständige Rechtssätze können bei der Falllösung in verschiedener Weise zusammenwirken. So können in einem Fall mehrere vollständige Rechtssätze gleichzeitig anwendbar sein. Derselbe Tatbestand kann in verschiedenen Rechtssätzen mit unterschiedlichen Rechtsfolgen verbunden werden (zum Beispiel kann eine Urkundenfälschung straf- und disziplinarrechtliche Folgen haben). Außerdem können durch ein und dasselbe Ereignis in einem Fall mehrere

Tatbestände gleichzeitig erfüllt sein. Zugleich mit Schadenersatz für einen materiellen Schaden kann dasselbe Ereignis zum Beispiel zu einem Anspruch auf Schmerzensgeld aufgrund eines gleichzeitig entstandenen immateriellen Schadens führen. Schließlich können verschiedene Normen mit verschiedenen in einem Fall erfüllten Tatbeständen dieselbe Rechtsfolge begründen. Etwa kann ein Schadenersatzanspruch im selben Fall zugleich aus Vertragsverletzung und unerlaubter Handlung entstehen, so dass der Schadenersatzanspruch überdeterminiert ist.

Neben solchen *kumulativen Normkonkurrenzen*, in denen mehrere Normen sich zu einer *komplexen Antwortnorm* ergänzen oder zumindest vereinbare Rechtsfolgen haben, kann es auch zu *Verdrängungssituationen* kommen, in denen verschiedene Normen für denselben Fall unvereinbare Rechtsfolgen festlegen. Solche Konflikte werden zum Teil durch explizite Festlegungen im Gesetz geregelt. Zudem existieren allgemeine Vorrangregeln. So bestimmt die generelle Normhierarchie unter anderem, dass Bundesrecht Landesrecht bricht und dass auf Bundes- wie Landesebene Verfassungsrecht über formellen Gesetzen steht. Weitere Vorrangregelungen führen zur Bevorzugung speziellerer gegenüber allgemeineren und neuerer gegenüber älteren Normen.

### 1.1.3 Auslegung und Subsumtion

Schon die Gewinnung der allgemeinen Norm, mit der in einem konkreten Fall zu arbeiten ist, ist also ein komplexer Vorgang. Auch wenn die relevanten Regelungen identifiziert und entsprechend ihren Bezügen zusammengefügt sind, können die meisten Fälle jedoch noch nicht durch einen direkten logischen Schluss gelöst werden. Es ist dann zwar bekannt, welche Tatbestandsmerkmale zu überprüfen sind. Das bedeutet aber in aller Regel noch nicht, dass ohne weiteres erkennbar ist, ob diese im konkreten Fall erfüllt sind oder nicht. Dies muss nicht an fehlender Information über den jeweiligen Sachverhalt liegen. Auch wenn alle relevanten Fakten über den Fall bekannt sind, ist oft nicht klar, wie diese sich zu den geforderten Tatbestandsmerkmalen verhalten.[5] Ob zum Beispiel die Behauptung gerechtfertigt ist, dass bestimmte Eigenschaften einer verschenkten Sache einen *Fehler* darstellen, dass jemand diesen wirklich *verschwiegen* und dies gar *arglistig* getan hat, hängt – neben der Sachlage im Fall – auch davon ab, wie diese Tatbestandsmerkmale genau zu verstehen sind. Tatbestandsmerkmale selber determinieren oft also nicht eindeutig eine Entscheidung. Sie bedürfen erst noch weiterer *Auslegung*.

[5] Vergleichbare Schwierigkeiten können sich auch für die *Rechtsfolgenseite* von Normen ergeben. Von dieser Problematik soll hier vereinfachend abgesehen werden.

#### (a) Auslegungskriterien

Die Frage, wie und unter welchen Gesichtspunkten diese Auslegung zu erfolgen hat, stellt ein zentrales Thema der juristischen Methodenlehre dar. Traditionell werden die folgenden – von Savigny (von Savigny (1840)) als *Sinnbestimmungsmittel* in die Auslegungslehre eingebrachten – vier Gesichtspunkte als zulässige auslegungsleitende Faktoren angesehen (die sogenannten *canones der Auslegung*):

- **Wortsinn.** Vom Wortsinn der Normformulierung ist – so klar er sich bestimmen lässt – bei der Auslegung auszugehen. Er markiert die Grenze, innerhalb derer sich mögliche Interpretationen zu halten haben (die sogenannte *Wortlautgrenze*).

- **Systematik.** Die Systematik der Normen schränkt den Auslegungsspielraum weiter ein. Zum Beispiel ist bei Normkonkurrenzen oft nur eine bestimmte Auslegungsvariante eines Tatbestandsmerkmals sinnvoll, da sonst – wie oben erläutert – eine Verdrängungssituation eintritt. Um Klarheit über den Sinn eines Merkmals in einer Norm zu erhalten, können außerdem zum Beispiel Vergleiche mit seiner Verwendung in verwandten Normen gezogen werden. Es kann auch anderen Normen, auf deren Anwendbarkeit sich ein Tatbestandsmerkmal in einer Hilfsnorm indirekt auswirkt, Information für die Auslegung des Merkmals zu entnehmen sein, oder es kann die Art und Weise betrachtet werden, in der ein Komplex von Normen einen Sachbereich insgesamt regelt.

- **Entstehungsgeschichte.** Auch historische Fakten zur Entstehung einer Norm können dazu beitragen, zu determinieren, wie diese auf konkrete Fälle zu beziehen ist. Hierzu gehört die Entstehungsgeschichte der Norm im engeren Sinne. Zu ihrer Ermittlung können beispielsweise Materialien herangezogen werden, die den Gesetzgebungsprozess dokumentieren. Darüber hinaus sind auch die gesellschaftlich-politischen Rahmenbedingungen zur Entstehungszeit der Norm und der damalige rechtliche Zusammenhang, in den sich die Norm ursprünglich einfügte, zu berücksichtigen. Schließlich spielt auch der Zweck eine Rolle, den der historische Gesetzgeber mit dem Erlass der Norm verfolgte.

- **Zweck.** Auch aus der Richtung der Ziele, die mit einer Regelung gegenwärtig zu erreichen sind, kann bestimmt werden, wie diese sinnvoll zu interpretieren ist. Als übergeordnetes Ziel der Rechtsordnung kann der bestmögliche gerechte Interessenausgleich in Konfliktfällen angesehen werde. Normen können daneben spezieller der Umsetzung bestimmter Prinzipien und Rechtsgrundsätze dienen (zum Beispiel der Wahrung

> der Rechtssicherheit oder dem Vertrauensschutz) oder auch enger eingegrenzte konkrete Zwecke haben. Vielen Normen kommt außerdem eine bestimmte Funktion hinsichtlich anderer Normen und Normkomplexe zu (so stützen und erhalten etwa strafrechtliche Bestimmungen gegen Eigentumsdelikte die Eigentumsordnung des bürgerlichen Rechts). Zweckerwägungen vermögen die Auslegung zu lenken, da in der Regel einige Interpretationen der Erfüllung solcher abstrakten und konkreteren Zielsetzungen eher dienlich sind als andere.

Savignys *canones* gehören zum methodologischen Kernbestand der Rechtswissenschaft und werden in dieser oder ähnlicher Weise in den meisten Standardwerken und Lehrbüchern an zentraler Stelle besprochen.[6] Vielfach wird jedoch zugleich kritisiert, sie seien unpräzise und nicht klar voneinander abgegrenzt. Zudem wird das Problem einer Rangordnung der einzelnen Kriterien diskutiert. Von Savigny selber scheint eine scharfe Abgrenzung oder Präzedenzordnung seiner *canones* nicht angenommen zu haben. Er weist darauf hin, die *canones* seien *nicht vier Arten der Auslegung, unter denen man nach Belieben wählen könnte, sondern verschiedene Tätigkeiten, die vereinigt werden müssen, wenn die Auslegung gelingen soll* (von Savigny (1840), 215).

Gegen die *canones* wird zudem eingewandt, sie erfassten die in der Auslegung tatsächlich relevante Information nur unvollständig. Insbesondere wird darauf verwiesen, dass neben dem auszulegenden Gesetz auch der bisherigen Rechtsprechung entnommenes "Richterrecht" entscheidungsbestimmend sein kann. Von unterschiedlichen Autoren wird der *gefestigten Rechtsprechung* eine rein heuristische Funktion (als "Erkenntnisquelle" Larenz (1991), 432), "präsumtive Verbindlichkeit" (Kriele (1967), 243 ff.) oder sogar der Status einer eigenständigen Rechtsquelle (zum Beispiel bei Fickentscher (1977), 143 f.) zuerkannt. Im Folgenden wird deutlich werden, dass jedenfalls *innerhalb* der an den *canones* orientierten Gesetzesauslegung regelmäßig Argumente durch Verweis auf Präjudizien gestützt werden.

#### (b) Deduktive Entscheidungsbegründung

In den meisten Fällen kann erst durch Auslegung entlang der eben beschriebenen Leitlinien die *Kluft zwischen Normformulierung und Sachverhaltsbeschreibung* (Koch und Rüßmann (1982)) überbrückt werden. Die relevanten Tatbestandsmerkmale müssen soweit präzisiert werden, dass ersichtlich ist, wie sie

[6] Zum Beispiel Larenz (1991), 339 ff. (wo noch der Gesichtspunkt *verfassungsgemäße Auslegung* hinzugenommen wird). Die hier gegebene Darstellung orientiert sich an Engisch (1997), Zippelius (2005) und Schwacke (2003).

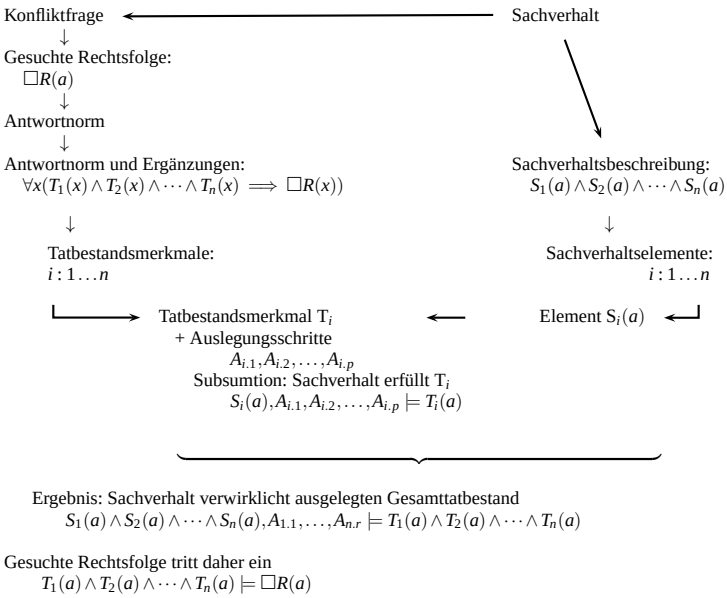



Abbildung 1.1: Subsumtion nach deduktivem Begründungsschema (vgl. Schwacke (2003), 61)

sich auf Aspekte des konkreten Sachverhalts beziehen. Erst dann kann der Sachverhalt unter den Tatbestand *subsumiert*, das heißt die Zuordnung einzelner Tatbestandsmerkmale zu Elementen des Sachverhalts tatsächlich vorgenommen werden. Gelingt dies, folgt die gesuchte Rechtsfolge aus der (wie oben dargestellt zusammengetragenen) allgemeinen Normformulierung, der Sachverhaltsbeschreibung *und* den interpretativen Annahmen. Schematisch lässt sich dieses Modell der Fallentscheidung darstellen wie in Abb. 1.1 (für den einfachen Fall eines konjunktiv strukturierten Tatbestands, der durch den Sachverhalt tatsächlich erfüllt wird).

Der Sachverhalt wirft eine Konfliktfrage auf, die wiederum die Suche nach einer bestimmten Rechtsfolge ($\Box R(a)$) motiviert. Anhand der gesuchten Rechtsfolge wird die Antwortnorm identifiziert und gegebenenfalls (z.B. durch zugehörige Hilfs- und Gegennormen) ergänzt. Erst dann ergibt sich ein

Gesamttatbestand, dem einzelne Tatbestandsmerkmale ($T_1, \ldots, T_n$) zu entnehmen sind. Für jedes Tatbestandsmerkmal ($T_i$) wird sodann untersucht, ob es im Sachverhalt verwirklicht ist. Wenn dies der Fall ist, lässt sich das instantiierte Tatbestandsmerkmal ($T_i(a)$) aus einem Bestandteil der Sachverhaltsbeschreibung ($S_i(a)$) und bestimmten zusätzlichen Prämissen ($A_{i.1}, \ldots, A_{i.r}$) ableiten (sog. *deduktives Nebenschema*). Diese zusätzlichen Prämissen werden durch Auslegung des jeweiligen Tatbestandsmerkmals unter den oben besprochenen Gesichtspunkten gewonnen. Ist die Ableitbarkeit der einzelnen instantiierten Tatbestandsmerkmale gezeigt, wird auch die instantiierte Rechtsfolge ableitbar (sog. *deduktives Hauptschema*).

Nach dem *Deduktivitätspostulat* (vgl. dazu Koch (2003)) ist eine schlüssige und vollständige Darstellung der Entscheidungsgründe entlang dieses Ableitungsschemas eine formale Minimalanforderung an methodengerechte Falllösungen. Insbesondere müssen sämtliche in der Auslegung gemachten interpretativen Annahmen mitsamt den sie stützenden Erwägungen in der Urteilsbegründung explizit gemacht werden, damit die Zuordnung zwischen Sachverhalt und Tatbestand Schritt für Schritt nachvollziehbar und überprüfbar ist.

Auch das *Subsumtionsmodell* der Rechtsanwendung als schrittweise Überprüfung und Feststellung von Tatbestandsmerkmalen stellt einerseits eine Leitidee der juristischen Methodologie dar und wird andererseits vielfach kritisiert. Besonders das deduktive Begründungsschema wird kontrovers diskutiert. Ausdrücklich anerkannt wird sein Wert von der *analytischen Methodenlehre* (Koch und Rüßmann (1982), Alexy (1996)). Sie betont die grundlegende Bedeutung der expliziten deduktiven Entscheidungsbegründung für die Rechtsstaatlichkeit und Transparenz der Rechtsprechung. Die *Strukturierende Rechtslehre* (Müller (1997), Müller (1994)) dagegen übt Fundamentalkritik an diesem Begründungsmodell. Sie stützt sich dabei vor allem auf zwei Argumente ähnlicher Stoßrichtung: (a) Es wird eingewandt, einzelnen sprachlichen Ausdrücken (generell und daher insbesondere in Normformulierungen) komme keine isolierbare Semantik zu; sie seien nur situiert, im Gesamttext und unter Berücksichtigung aller intertextuellen Bezüge sinnvoll zu interpretieren. (b) Es wird argumentiert, die Bedeutung sprachlicher Ausdrücke sei grundsätzlich nicht vorgegeben, sondern vollständig in der Anwendung erzeugt. *Bedeutung und Wert der gebrauchten Zeichen werden in den realen sprachlichen Handlungen nicht nur nicht herausgefunden beziehungsweise nur modifiziert, sondern (...) überhaupt erst geschaffen* (Müller (1994), 377). Normen seien deshalb dem Normtext nicht zu entnehmen, sondern würden erst in einem Prozess der "Rechtsarbeit" (für die der Text allerdings ein "Eingangsdatum" darstelle) hervorgebracht. Aufgrund dieser Argumente wird die in der klassischen Auslegungslehre geforderte Bindung der Auslegung an die wörtliche Bedeutung einzelner Tatbestandsmerkmale als ebenso obsolet angesehen wie die Aufglie-

derung der Entscheidungsbegründung in einzelne, isoliert interpretierte Prämissen und Konklusionen.

Es erscheint schwer vorstellbar, dass auf der Grundlage eines so radikal holistischen und pragmatischen Ansatzes eine Sprachtheorie aufgebaut werden kann. Hierfür dürfte auf die Annahme eines Kompositionalitätsgrundsatzes kaum zu verzichten sein, nach dem die Vielfalt möglicher Bedeutungen aus einem endlichen Grundinventar zumindest prinzipiell entlang der Struktur sprachlicher Ausdrücke systematisch konstruierbar ist. Es stellt sich zudem die Frage, wie Kommunikation gelingen kann, wenn Bedeutung im Sprachgebrauch permanent *vollständig neu* konstituiert wird. Ebenso ist auf der Basis dieser Annahme nicht erklärbar, wieso Ausdrücke der Sprache (abgesehen von syntaktischen Kriterien) richtig und falsch verwendet werden können.

Zwar werden gewisse zentrale Begriffe der Rechtsordnung durch ein komplexes, nicht klar abgrenzbares Normengeflecht bestimmt (vergleiche hierzu 1.2.2). Daraus kann aber nicht verallgemeinernd geschlossen werden, Rechtsbegriffe seien generell semantisch überhaupt nicht isoliert und über einzelne Anwendungsfälle hinaus bestimmbar. Wie sich im Rahmen dieser Arbeit zeigen wird, geht jedenfalls die Praxis der Gesetzgebung und Rechtsprechung offenbar ohne weiteres davon aus, dass sich über die wörtliche Bedeutung einzelner isolierter Tatbestandsmerkmale sinnvoll reden lässt.

Ein weniger grundlegender Einwand gegen das deduktive Begründungsmodell macht geltend, die zur Rechtfertigung praktischer Schlussfolgerungen benötigte Vielfalt und Komplexität juristischer Schlussweisen lasse sich nicht adäquat im Rahmen einer logischen Rekonstruktion erfassen (so z.B. Neumann (2004), 314). In der vorliegenden Arbeit kann auf diese Diskussion nicht genauer eingegangen werden. Es soll nur angemerkt werden, dass das deduktive Begründungsschema formale Anforderungen an eine ideal explizite Darstellung der Argumentation formuliert. Diese Argumentation muss aber natürlich irgendwo ihren Anfang nehmen. Letztbegründungen für wertende Prämissen – die in der Tat nicht allein logisch herzuleiten sind – erfordert das deduktive Begründungsmodell nicht (darauf weist z.B. Alexy (1996), u.a. 224 ff., 233 ff., hin.). Ebenso wenig reklamiert es für sich, die Funktion aller Sätze in juristischen Argumentationen vollständig beschreiben zu können. In diesem Kapitel wird auf die Vielfalt spezifischer Funktionen einzelner Aussagen in Urteilsbegründungen noch genauer einzugehen sein (siehe 1.3).

Festzuhalten bleibt somit bis hierher folgende, in dieser Form wohl nicht nur mit der analytischen Position vereinbare Erkenntnis: Für die Rechtsanwendung ist von zentraler Bedeutung, wie Tatbestandsbeschreibungen *ihrer sprachlichen Bedeutung* nach auf Sachverhalte bezogen werden. Dies geschieht nach einem *methodologisch reflektierten*, *rationalen* und *stark normierten Verfahren*, dessen Schritte für eine akzeptable Entscheidungsbegründung *in logisch schlüssi-*

*ger Weise explizit* gemacht werden müssen. Der Strukturierenden Rechtslehre kann dabei zugestanden werden, dass Kontext und Sprachgebrauch konstitutive Faktoren der sprachlichen Bedeutung sind, deren Beitrag in der Argumentation zu berücksichtigen und gegebenenfalls ebenfalls explizit darzustellen ist.

## 1.2 Definitionen in Gesetzgebung und Rechtsprechung

Unsere Betrachtung des Auslegungsprozesses hat ergeben, dass dieser sich in hohem Maße auf semantische Erwägungen stützt. Die Bedeutung der Tatbestandsformulierung einer Norm stellt zugleich den Ausgangspunkt und die Grenze für ihre Anwendung dar. Wir möchten nun untersuchen, auf welche Weise dieser allgemeine Grundsatz in der methodengerechten Rechtsprechung konkret umgesetzt wird. Unser Ausgangspunkt ist dabei die Frage: Wie wird die Bedeutung der in Normtexten verwendeten Ausdrücke bestimmt?

Der aus dem Rechtsstaatsprinzip des Grundgesetzes hergeleitete Bestimmtheitsgrundsatz[7] besagt, dass Tragweite und Anwendungsbereich von Normen für den Adressaten (d.h. in vielen Fällen den nicht mit der juristischen Fachsprache vertrauten Bürger) erkennbar sein müssen. Dies ist nur möglich, wenn die Rechtssprache sich zumindest bis zu einem gewissen Grad am generellen Sprachgebrauch orientiert. Entsprechend bestimmt beispielsweise § 42 Abs. 5 der *Gemeinsamen Geschäftsordnung der Bundesministerien*:

> Gesetze müssen sprachlich einwandfrei und sollten so weit wie möglich für jedermann verständlich gefasst sein.

Zugleich begründet der Bestimmtheitsgrundsatz auch die Forderung nach einem einheitlichen und möglichst exakten Sprachgebrauch in Gesetzgebung und Rechtsprechung. Tragweite und Anwendungsbereich von Normen sind nur dann erkennbar, wenn der *Inhalt der Normformulierung* hinreichend klar und eindeutig erkennbar ist.

Die Bedeutung alltagssprachlicher Ausdrücke ist jedoch in der Regel nicht erschöpfend und endgültig bestimmt. Im Alltag stellen Uneinigkeit über die Verwendungsweise von Begriffen oder unzureichende Bestimmtheit allerdings nur selten ein Kommunikationshindernis dar. Das notwendige Maß an Übereinstimmung und der kommunikativ erforderliche Bestimmtheitsgrad können meist ohne weiteres im Verlauf der Kommunikation selber erzielt werden. Diese Flexibilität ist eine Grundvoraussetzung für die Verwendbarkeit der natürlichen Sprache als universelles Kommunikationsmedium. Sie führt jedoch auch

[7] Für das Strafrecht ist der Bestimmtheitsgrundsatz in Artikel 103 II GG explizit formuliert.

dazu, dass den Anforderungen des Bestimmtheitsgrundsatzes durch eine alleinige Orientierung der Rechtssprache an der Alltagssprache nicht entsprochen werden kann.

Daher wird in der Gesetzgebung ein Kompromiss zwischen *Allgemeinverständlichkeit* und *Fachsprachlichkeit* gesucht, der in vielen Fällen (das wohl bekannteste Beispiel ist die rechtssprachliche Unterscheidung zwischen den normalsprachlich äquivalenten Begriffen *Besitz* und *Eigentum*) vom alltäglichen Sprachgebrauch abweicht. Wie in allen Fachsprachen erhalten auch in der Rechtssprache wichtige Termini einen fachspezifischen Sinn, der durch *Definitionen* festgelegt wird.

Während jedoch in den Naturwissenschaften und vielen Bereichen der Technik durch Definitionen ein geschlossenes System (mehr oder weniger) zweifelsfrei entscheidbarer Begrifflichkeiten aufgebaut wird, ist dies in der Rechtssprache prinzipiell nicht möglich. Die gesetzlich festgeschriebenen *Legaldefinitionen* liefern nur ein Grundgerüst, das in der Rechtsprechung weiter ausgearbeitet werden muss. Im Folgenden betrachten wir den Prozess der juristischen Begriffsbestimmung näher. Zu diesem Zweck gehen wir zunächst kurz auf ausgewählte Aspekte aus der philosophisch-wissenschaftstheoretischen Diskussion zu Definitionen sowie auf einige moderne, praxisorientierte Festlegungen in den sogenannten terminologischen Grundsatznormen (DIN 2330, 2331 und 2342) zur Festlegung technischer Terminologie ein. Vor diesem Hintergrund untersuchen wir dann genauer die Besonderheiten bei der Bestimmung rechtssprachlicher Begrifflichkeiten.

### 1.2.1 Terminologieaufbau durch Definitionen

Die Beschäftigung mit Definitionen hat eine lange philosophische Tradition. Schon in der klassischen Philosophie wurden zwei zentrale Anforderungen an Definitionen formuliert, die sich folgendermaßen zusammenfassen lassen:

**Äquivalenz:** Eine Definition soll den zu definierenden Begriff (das *Definiens*) durch einen *äquivalenten* Ausdruck (das *Definiendum*) bestimmen.

**Essentialität:** Eine Definition soll die *wesentlichen* Eigenschaften erfassen, kraft derer Gegenstände unter einen Begriff fallen.

Die Äquivalenzbedingung fordert, dass eine Definition notwendige und (gemeinsam) hinreichende Bedingungen für die Anwendung des Definiens angibt. Sie sichert damit die Trennschärfe der Definition. Alle und nur die Gegenstände, die unter das Definiens fallen, sollen auch durch das Definiendum bezeichnet werden. Die Essentialitätsbedingung grenzt den Inhalt des Definiens enger

ein, denn nicht alle möglichen zu einem Begriff äquivalenten Beschreibungen benennen für den Begriff wesentliche Eigenschaften.[8]

Die philosophische Diskussion zum Thema Definitionen befasste sich vor allem mit Grundlagenfragen, die an diese beiden Bedingungen anknüpfen. So wurde diskutiert, ob die Erfüllung der Essentialitätsbedingung nur durch ontologische Erkenntnisse über das Wesen von Dingen zu erreichen sei, ob sie durch empirische Analyse des Sprachgebrauchs erzielt werden solle oder ob Definitionen bloße Bennenungskonventionen darstellen und es sich bei dem Kriterium somit eigentlich um ein mit Definitionen verbundenes Postulat handle.[9] Mit der Entwicklung der technischen Disziplinen in der jüngeren Vergangenheit stellte sich jedoch in den verschiedensten Fachrichtungen die praktische Aufgabe, große Terminologiebestände zu normieren. Damit entstand ein Bedarf nach konkreten, praktisch umsetzbaren Standards für eine methodengeleitete Terminologiearbeit. Solche Standards enthalten die terminologischen Grundsatznormen, in Deutschland die DIN-Normen 2330 (*Begriffe und Benennungen - Allgemeine Grundsätze*), 2331 (*Begriffssysteme und ihre Darstellung*) und 2342 Teil 1 (Begriffe der Terminologielehre - Grundbegriffe).

Zum Definitionsbegriff hält DIN 2330 unter anderem Folgendes fest:

> Beim Definieren wird ein Begriff mit Hilfe des Bezugs auf andere Begriffe innerhalb eines Begriffssystems festgelegt und beschrieben und damit gegenüber anderen Begriffen abgegrenzt. Die Definition bildet die Grundlage für die Zuordnung einer Benennung zu einem Begriff; ohne sie ist es nicht möglich, einem Begriff eine geeignete Benennung zuzuordnen. (...) Sofern in Definitionen Begriffe nicht allgemein bekannt sind, müssen sie an gleicher oder anderer, genau bezeichneter Stelle definiert sein.

Zu den grundsätzlichen Fragestellungen, die in der Philosophie im Zusammenhang mit Definitionen diskutiert wurden, trifft DIN 2330 pragmatische Festlegungen. Die Essentialitätsbedingung wird folgendermaßen formuliert:

---

[8] Uns ist bewusst, dass wir hier "im Nebensatz" eine weitreichende Annahme bezüglich einer zentralen sprachphilosophischen Frage formulieren. Die notwendige Grundsatzdiskussion zum Bedeutungsbegriff wäre im Rahmen dieses Kapitels allerdings nicht ansatzweise zu leisten. Ein Leser, der unsere Explikation der Essentialitätsbedingung aus fundamentalen theoretischen Bedenken ablehnt, dürfte aufgrund vergleichbarer Erwägungen in unserer Arbeit auch andernorts Einspruch erheben (oder schon erhoben haben).

[9] Den ersten, *essentialistischen* Ansatz können – natürlich unbeschadet ansonsten bestehender großer Unterschiede – beispielsweise Plato, Aristoteles, Kant und Husserl zugerechnet werden. Den zweiten, *linguistischen* Ansatz verfolgen (es gilt derselbe Vorbehalt) Mill, Moore und Robinson (Robinson (1964)), den dritten, *präskriptiven* schließlich Pascal, Russell, Quine, Goodman und Carnap – vgl. Abelson (1967).

> In Abhängigkeit von der Zielrichtung der jeweiligen Terminologiearbeit sind unterschiedliche Eigenschaften für die Begriffsbildung und -abgrenzung als wesentlich heranzuziehen.

Sowie:

> Die in eine Definition aufzunehmenden Merkmale müssen für das jeweilige Fachgebiet wesentlich sein. Die Merkmale müssen deshalb so gewählt werden, daß sie die Einordnung in das entsprechende Begriffssystem ermöglichen.

Definitionen sollen nach dieser Bestimmung ein strukturiertes und möglichst vollständiges terminologisches System aufbauen. Unbestimmt dürfen nur Begriffe bleiben, die als bekannt vorausgesetzt werden können. Inhalt und Umfang von Definitionen bestimmen sich im Sinne der zitierten Feststellungen jedoch nicht nach ontologischen oder sprachtheoretischen Kriterien.

Insbesondere das Essentialitätskriterium wird zu fachspezifischen Gegebenheiten relativiert. Zentral sind die Anforderungen des jeweiligen Fachs und insbesondere der Gesichtspunkt der Systematisierung der gesamten Terminologie. Dabei werden drei Voraussetzungen gemacht: (1) Die terminologisch zu regelnde Fachsprache lässt sich klar von der unproblematisch verwendbaren Gemeinsprache abgrenzen. (2) Die verwendeten Begriffe lassen sich durch Zerlegung in Merkmale abschließend bestimmen, und (3) sie fügen sich zu einem System zusammen. DIN 2330 (und ebenso die weiteren terminologischen Normen) zielt in erster Linie auf die Anleitung der Terminologiearbeit im naturwissenschaftlich-technischen Bereich. Ob die genannten Voraussetzungen für die Rechtssprache als erfüllt angesehen werden können, ist im Folgenden zu überprüfen.

### 1.2.2 Besonderheiten der Rechtssprache

Auch in Gesetzen werden zentrale Begriffe systematisch durch Definitionen bestimmt. So schreiben beispielsweise §§ 13 und 14 BGB fest:

> § 13 Verbraucher ist jede natürliche Person, die ein Rechtsgeschäft zu einem Zwecke abschließt, der weder ihrer gewerblichen noch ihrer selbständigen beruflichen Tätigkeit zugerechnet werden kann.
>
> § 14 (1) Unternehmer ist eine natürliche oder juristische Person oder eine rechtsfähige Personengesellschaft, die bei Abschluss eines Rechtsgeschäfts in Ausübung ihrer gewerblichen oder selbständigen beruflichen Tätigkeit handelt.

§ 13 bestimmt den Begriff *Verbraucher*, indem der übergeordnete Begriff *natürliche Person* durch das Merkmal *Abschluss eines Rechtsgeschäfts* und das (von diesem abhängige) Merkmal *kein beruflicher oder gewerblicher Zweck* determiniert wird. Der Begriff *Verbraucher* wird damit dem Begriff *natürliche Person* untergeordnet. Zugleich grenzt das letztgenannte Merkmal den Begriff *Verbraucher* auch von dem in § 14 definierten Begriff *Unternehmer* ab. Dieser umfasst nach § 14 wiederum neben natürlichen auch juristische Personen. Die beiden zitierten Definitionen systematisieren also terminologisch den Bereich der Beteiligten an Rechtsgeschäften. Dabei wird Bezug genommen auf eine generelle Systematisierung möglicher Handlungsbeteiligter (die im BGB zu Grunde gelegte Einteilung in Sachen und Personen sowie natürliche und juristische Personen) sowie wirtschaftlicher Tätigkeiten (selbständig-beruflich bzw. gewerblich oder nicht).

Die zitierten Definitionen entsprechen somit in Funktion und Aufbau den im vorigen Abschnitt diskutierten Festlegungen der terminologischen Grundsatznormen. Solche in Normtexten enthaltenen terminologischen Definitionen werden als *Legaldefinitionen* bezeichnet. In neueren Gesetzen sind die relevanten Legaldefinitionen oft zusammengefasst am Anfang des Gesetzestextes bzw. einzelner Abschnitte aufgelistet und damit explizit als terminologische Festlegungen vom eigentlichen Regelungsgehalt des Gesetzes abgegrenzt.

Die Abdeckung durch Legaldefinitionen beschränkt sich jedoch in allen Rechtsbereichen auf eine relativ geringe Zahl zentraler Begriffe. Ansonsten ist – anders als im naturwissenschaftlich-technischen Bereich – schon die Abgrenzung eines rechtlichen Fachvokabulars gegenüber der Gemeinsprache und auch gegenüber anderen Fachsprachen mit Schwierigkeiten verbunden. Die gesellschaftliche Realität, auf die in der Rechtsprechung Bezug genommen wird, ist im Ganzen nicht überschaubar und permanentem Wandel unterworfen. Es ist daher nicht im Vorhinein absehbar, welche Begriffe in Zukunft rechtliche Relevanz erhalten werden. Im Gegensatz zu anderen Fachsprachen hat die rechtssprachliche Terminologie immer wieder Begriffe aus verschiedensten anderen Bereichen zu assimilieren (die oben zitierten Legaldefinitionen §§ 13 und 14 BGB etwa fanden erst im Jahr 2000 Eingang in den Gesetzestext).

Darüber hinaus bestehen zwischen der Rechtssprache und den Fachsprachen der Naturwissenschaften und technischen Disziplinen aber auch grundsätzliche Unterschiede auf semantischer und pragmatischer Ebene. Einige für die Rechtssprache wichtige *Begriffstypen* entziehen sich aus prinzipiellen Gründen einer abschließenden terminologischen Systematisierung durch Legaldefinitionen. Wir gehen im Folgenden näher auf die juristisch besonders relevanten *institutionellen* und *evaluativen Begriffe* ein. Selbst legaldefinierte Begriffe sind zudem – wie wir im Anschluss erläutern werden – in aller Regel in ihrer Anwendbarkeit nicht umfassend festgelegt.

### (a) Institutionelle Begriffe

Der Diskurs der Naturwissenschaft und Technik zielt auf die Vermittlung von Information über natürliche Tatsachen. Entsprechend ist die zentrale Funktion der verwendeten Begriffe die Referenz auf sprachunabhängige Entitäten. Hauptaufgabe des Rechtssystems ist dagegen die Ordnung menschlichen Handelns. Hierfür spielt der Bezug auf *institutionelle Tatsachen*, die erst durch die normative Wirkung der Rechtsordnung entstehen, eine zentrale Rolle.[10]

Als Bestandteile von Rechtsnormen sind alle Rechtsbegriffe in gewissem Sinne normbezogen: Ihre Verwendung dient neben der Referenzfunktion der Begründung (und, bei Begriffen auf der Rechtsfolgenseite, der Ausprägung) von Rechtsfolgen und hat somit normative Konsequenzen. Bestimmte zentrale Begriffe der Rechtsordnung, wie *Eigentum*, *Besitz* oder *Ehe*, nehmen jedoch (in ihrem rechtssprachlichen Sinn) überhaupt nicht unmittelbar deskriptiv auf empirisch Gegebenes Bezug. Die Anwendbarkeit solcher Begriffe setzt die Existenz bestimmter menschlicher, speziell gewisser rechtlicher *Institutionen* im Sinne von Systemen konstitutiver Regeln voraus. Sie verweisen auf ganze Komplexe von Vorschriften über das Zustandekommen, die Beendigung und die Folgen (z.B. Ansprüche und Pflichten) bestimmter Tatsachen innerhalb der Rechtsordnung.

Der Begriff *Besitz* wird beispielsweise im Kern durch §§ 854 und 856 BGB (Erwerb bzw. Beendigung des Besitzes) bestimmt. Verschiedene weitere Normen in Buch 3, Abschnitt 1 BGB ergänzen diesen Kerninhalt, indem sie Genaueres über Erwerb bzw. Beendigung des Besitzes festlegen und Grundsätzliches zu den an das Bestehen von Besitz geknüpften Rechtsfolgen regeln. Auch Regelungen aus anderen Gesetzen befassen sich aber noch mit Folgen des Bestehens von Besitz, wobei der Übergang von definitorischer zu nicht definitorischer Information fließend ist. § 3 Abs. 4 KrW-/AbfG etwa legt fest, dass der Besitz einer Sache unter bestimmten Bedingungen (u.a. *Unmöglichkeit der Verwendung gemäß dem ursprünglichen Zweck* und *Gefährdungspotential für die Allgemeinheit*) die Verpflichtung nach sich zieht, sich der Sache zu entledigen. So wie im Falle des Begriffs *Besitz* ist auch das mit anderen institutionsbezogenen Begriffen verknüpfte Normengeflecht nicht ohne weiteres abzugrenzen. Ähnlich wie die in der Wissenschaftstheorie diskutierten *theoretischen Begriffe* (vgl. u.a. Carnap (1956), Hempel (1950), Quine (1951)) sind solche Begriffe daher nicht durch eine abgeschlossene Aufzählung von Merkmalen definierbar. Wie bereits oben (1.1.3) angesprochen zwingt diese Tatsache aber nicht

[10] Die Unterscheidung zwischen natürlichen und institutionellen Tatsachen geht zurück auf Anscombe (1958) und Searle (1969), 50 ff.; die Rolle institutioneller Tatsachen für die Rechtsordnung untersuchen zum Beispiel Ruiter (1993) und MacCormick und Weinberger (1986)

zur Annahme der bedeutungsskeptischen Extremposition, über die Bedeutung isolierter Rechtsbegriffe seien generell nie sinnvolle Angaben möglich.

### (b) Evaluative Begriffe

Die Rechtfertigung normativer Konsequenzen beruht letztlich auf wertenden Entscheidungen. Für die Rechtsordnung sind grundlegende Wertentscheidungen zu einem großen Teil durch den Gesetzgeber vorweggenommen. In vielen Fällen verbleiben jedoch – vor allem hinsichtlich detaillierterer Bewertungen – in der Rechtsanwendung Spielräume. Dies äußert sich darin, dass in vielen Normen Tatbestandsmerkmale unter Verwendung *evaluativer Begriffe* ausgedrückt werden. Beispiele sind etwa die Rechtsbegriffe *gute Sitten* (z.B. §§ 138 Abs. 1, 826 BGB), *wichtiger Grund* (z.B. §§ 314, 626 BGB) oder *angemessene Vergütung* (im Urheberrechtsgesetz). Aber auch für die Anwendung von Konzepten, die im alltagssprachlichen Verständnis nicht mit Wertvorstellungen im engeren Sinne verknüpft sind, sind oft wertende (allerdings moralisch relativ neutrale) Entscheidungen grundlegend. Ein Beispiel hierfür ist der Begriff *Stand der Technik* im Umweltrecht.

Auch evaluative Begriffe sind durch Definitionen nicht abschließend bestimmbar, da ihre deskriptiv-referentielle Komponente nicht allgemein festgelegt werden kann. Ihre Verwendung erfordert vom Sprecher eine Wertung. Deskriptive Kriterien können nur jeweils in Abhängigkeit von dieser Wertung angegeben werden.[11] In der juristischen Methodenlehre werden solche Begriffe deshalb auch als *wertausfüllungbedürftig* bezeichnet.

Für wertausfüllungsbedürftige Begriffe legt der Gesetzgeber häufig anstatt deskriptiver Merkmale prozedurale Kriterien und leitende Gesichtspunkte für die Vorgehensweise beim Auffinden der betreffenden Wertung fest. Solche Angaben können knapp und sehr allgemeiner Natur sein. So enthält beispielsweise § 314 BGB (und in ähnlicher Formulierung auch § 626 BGB) den Hinweis, die Entscheidung über das Vorliegen eines *wichtigen Grundes* habe *unter Berücksichtigung aller Umstände des Einzelfalls und unter Abwägung der beidersei-*

[11] Die Unterscheidung einer evaluativen und einer deskriptiven Bedeutungskomponente bei evaluativen Begriffen geht auf Hare (1952) zurück. Dort wird angenommen, dass ein Sprecher mit solchen Begriffen Gegenstände oder Handlungen aufgrund seiner Wertung bestimmter deskriptiver Eigenschaften *empfiehlt*, und sich damit zugleich verpflichtet, dieselbe Empfehlung immer bei Vorliegen derselben deskriptiven Merkmale entsprechend auszusprechen. Hat sich eine bestimmte Wertung fest etabliert, kann es vorkommen, dass bei einem ursprünglich evaluativen Begriff die deskriptive Komponente die Bedeutung voll ausschöpft (Hare (1952), 120 ff.). Auch in der Rechtssprache sind in einigen Fällen für eigentlich (d.h. dem gemeinsprachlichen Sinn nach) evaluative Begriffe durch den Gesetzgeber deskriptive Kriterien auf der Basis vorgegebener, etablierter Wertungen festgeschrieben (vgl. für konkrete Beispiele hierzu Schwacke (2003), 21 f.)

*tigen Interessen* zu erfolgen. Es kann sich aber auch um detaillierte und relativ konkrete Kriterien handeln. Zum Begriff *Stand der Technik* legt etwa das KrW-/AbfG[12] in Anhang III Folgendes fest:

> Bei der Bestimmung des Standes der Technik sind unter Berücksichtigung der Verhältnismäßigkeit zwischen Aufwand und Nutzen möglicher Maßnahmen sowie des Grundsatzes der Vorsorge und der Vorbeugung, jeweils bezogen auf Anlagen einer bestimmten Art, insbesondere folgende Kriterien zu berücksichtigen:
>
> [*Es folgt eine Liste zwölf konkreter Kriterien, unter anderem:*]
>
> - Einsatz abfallarmer Technologie
> - Förderung der Rückgewinnung und Wiederverwertung der bei den einzelnen Verfahren erzeugten und verwendeten Stoffe und gegebenenfalls der Abfälle
> - Notwendigkeit, Unfällen vorzubeugen und deren Folgen für den Menschen und die Umwelt zu verringern

Selbst solche ausführlichen Anweisungen liefern jedoch keine Definitionen im oben dargestellten Sinne. Weder leisten sie eine systematische Einordnung des betreffenden evaluativen Begriffs, noch erfüllen sie das Äquivalenz- und Essentialitätskriterium. Grundlage der Begriffsverwendung bleibt eine im jeweiligen Einzelfall vorzunehmende Wertung. Die gesetzliche Begriffsbestimmung dient dazu, diese Wertung hinsichtlich festgelegter Gesichtspunkte auf Richtigkeit überprüfbar zu machen und damit in einem gewissen Umfang zu binden.

#### (c) Prinzipielle Unabgeschlossenheit

Viele – zum Teil zentrale – Rechtsbegriffe können also aufgrund ihrer integralen Einbindung in das Rechtsgefüge oder ihrer Wertbezogenheit nicht durch kompakte, abschließende Definitionen bestimmt werden. Selbst abgesehen von solchen prinzipiell undefinierbaren Rechtsbegriffen kann für die Rechtssprache jedoch kein abgeschlossenes terminologisches System aufgebaut werden. Auch legaldefinierte Gesetzesbegriffe sind effektiv nur selten abschließend und umfassend bestimmt. Bezüglich gesellschaftlicher Phänomene existiert keine der naturwissenschaftlichen Beobachtung vergleichbare objektive Perspektive.

Durch Definitionen können Rechtsbegriffe daher kaum einmal[13] auf zweifelsfrei (also z.B. durch Messung oder operationalisierte Kriterien) feststellbare

[12] Der Begriff ist hier relevant, um Anforderungen an die Abfallbeseitigung festzulegen.

[13] Abgesehen wohl v.a. von Zahlbegriffen für Zeit- und Entfernungsangaben oder Geldbeträge, vgl. z.B. Engisch (1997), 138 ff.

Grundbegriffe zurückgeführt werden. Ab einem bestimmten Punkt muss für die juristische Begriffsbestimmung auf undefinierte Ausdrücke der Alltagssprache zurückgegriffen werden. So legt zum Beispiel § 7 UrhG den Terminus *Urheber* mittels der Begriffe *Schöpfer* und *Werk* wie folgt fest:

> Urheber ist der Schöpfer des Werkes.

Jedoch bleibt im UrhG unbestimmt, wann eine Person im Zusammenhang dieses Paragraphen als *Schöpfer eines Werks* gelten soll.

Ergänzende Legaldefinitionen verschieben zwar den Übergang zur Anwendung gemeinsprachlicher Begriffe (und den damit einhergehenden Unbestimmtheitsproblemen). Für den Begriff *Werk* etwa ist in § 2 UrhG zwar folgende notwendige Bedingung angegeben:

> Werke im Sinne dieses Gesetzes sind nur persönliche geistige Schöpfungen.

Die drei hier im Definiens genannten Begriffe *persönlich*, *geistig* und *Schöpfung* allerdings sind im Gesetz nicht näher definiert. Über die Anwendbarkeit dieser Ausdrücke muss also jeweils durch Auslegung entschieden werden. Hierfür ist zunächst von ihrem alltagssprachlichen Sinn auszugehen. Jedem der Begriffe haftet damit ein gewisses Maß an Unbestimmtheit an, das sich auch auf die Verwendung des Urheberbegriffs überträgt.

Häufig sind alltagssprachliche Begriffe *randbereichsunscharf*.[14] Das heißt, obwohl eine große Zahl von Fällen klar als Positiv- oder Negativexemplare des Begriffs klassifiziert werden können, existiert eine selbst nicht klar umgrenzbare Menge von Fällen, deren Ähnlichkeit mit Positiv- *und* Negativbeispielen beide Klassifikationen zu rechtfertigen scheint. Beispielsweise ist bei an Vorbildern orientierten Produkten der angewandten Kunst oft strittig, ob sie noch als Schöpfungen und damit Werke im Sinne der zitierten Definition anzusehen sind, oder ob nur eine Einordnung als gewöhnliche handwerkliche Erzeugnisse gerechtfertigt ist (z.B. im Fall von Schmuck nach traditionellen Mustern, OLG München 29. Zivilsenat, 25. Februar 1993, 29 U 2918/92 und BGH 1. Zivilsenat, 22. Juni 1995, I ZR 119/93).

Die sichere Verwendung alltagssprachlicher Begriffe ist zudem an eine Vielzahl impliziter und selbst nur teilweise bestimmter Voraussetzungen und Rahmenbedingungen geknüpft.[15] In Fällen, in denen solche impliziten Vorbedingungen der Anwendbarkeit verletzt sind, kann der alltägliche Sprachgebrauch

[14] Zu einer genauen Klassifikation natürlich-sprachlicher Unbestimmtheitsphänomene vgl. Pinkal (1980).

[15] Zur Bezeichnung der aus dieser Tatsache resultierenden *potentiellen Vagheit* natürlich-sprachlicher Begriffe spricht Waismann (1945) metaphorisch von deren *Porosität* (engl. *open texture*)

jede Orientierungswirkung für die Verwendung eines Begriffs in der Rechtssprache verlieren. Bei Empfang eines religiösen Werkes durch Diktat aus dem Jenseits (KG Berlin 5. Zivilsenat, 27.03.1992, 5 U 6695/91) ist zum Beispiel völlig unklar, ob nach dem gemeinsprachlichen Sinn von einer *persönlichen geistigen Schöpfung* des Empfängers gesprochen werden kann. Aber auch in gewöhnlicheren Fällen können ähnlich gelagerte Unklarheiten bezüglich des Werkbegriffs nach § 2 UrhG entstehen, so zum Beispiel, wenn Werke gemeinschaftlich produziert wurden (z.B. von einem Architektenteam, Hanseatisches Oberlandesgericht Hamburg 5. Zivilsenat, 05.07.2006, 5 U 105/04) oder zunächst nicht als Kunstwerk, sondern als Gebrauchsgegenstand konzipiert wurden (z.B. Bauhaus-Hocker, OLG Düsseldorf 20. Zivilsenat, 30.05.2002, 20 U 81/01).

Die Rechtsordnung gibt den Gerichten jedoch einen Entscheidungszwang vor. Daher muss die Anwendbarkeit von Normen auch in solche Fällen geprüft und entschieden werden. Es müssen dann im Rahmen der Auslegungskriterien zulässige fallentscheidende Präzisierungen vorgenommen werden. Im oben angesprochenen Fall des Bauhaus-Hockers, den der Urheber Marcel Breuer selber nicht als kreativ-künstlerische persönliche Schöpfung angesehen hatte, argumentierte das OLG Düsseldorf folgendermaßen:

> Die Frage, ob eine persönlich-geistige Schöpfung im Sinne von § 2 Abs. 2 UrhG vorliegt, bemisst sich auch im Hinblick auf die schutzwürdigen Interessen der Erben des Urhebers nach objektiven Kriterien, nämlich nach der Anschauung der Verkehrskreise zur Zeit der Schöpfung des Werkes.

Je nach Bestimmtheitsgrad eröffnen sich für einzelne Tatbestandsmerkmale unterschiedlich große Auslegungsspielräume. Der Bestimmtheitsgrad von Rechtsbegriffen kann somit immense praktische Konsequenzen für die Rechtsunterworfenen haben. Gelingt es in einem rechtlichen Konflikt einer Partei, zu zeigen, dass die gegnerische Argumentation auf Begriffsbestimmungen beruht, die den aktuellen Fall nicht eindeutig erfassen, eröffnet sie sich damit die Möglichkeit, weitere Argumente für die eigene Position geltend zu machen. Im eben besprochenen Beispiel etwa konnte aufgrund der Vagheit der gesetzlichen Formulierung *persönliche geistige Schöpfung* der Gesichtspunkt *Schutz der Interessen der Erben* in die Erwägungen des Gerichts einbezogen werden. Es liegt daher in der Regel im Interesse mindestens einer der Konfliktparteien, die Trennschärfe der einschlägigen gesetzlichen Definitionen im konkreten Fall genau zu überprüfen. Praktisch werden Vagheitsbereiche und begriffliche Unschärfen so in der juristischen Argumentation systematisch ausgelotet.

Unbestimmtheit wird in der Rechtssprache durch Legaldefinitionen also nicht ausgeräumt. Die frei bleibenden Bedeutungsspielräume der Tatbestands-

merkmale eröffnen jedoch die Möglichkeit, durch fallweise Präzisierung einzelfallgerechte Entscheidungen unter Wahrung der allgemeinen Gesetzesbindung zu begründen. Das Präzisierungspotential alltagssprachlicher Ausdrücke im Gesetzestext garantiert insbesondere, dass einmal fixierte Normen auch auf lange Sicht flexibel genug bleiben, um eine sinnvolle Anwendung in immer neuen konkreten Fällen zu erlauben. Bei näherer Betrachtung erweist sich somit das trotz Legaldefinitionen verbleibende Maß an begrifflicher Offenheit als unerlässlich für das Funktionieren des Rechtssystems.[16]

## 1.3 Untersuchung definitionsbasierter Argumentationen

Im Regelfall können durch Legaldefinitionen festgeschriebene Rechtsbegriffe erst durch weitere Auslegung auf konkrete Sachverhalte bezogen werden. Auch im Rahmen der Auslegung müssen dafür begriffliche Festlegungen getroffen werden. Der in Normtexten definierte "terminologische Grundbestand" wird somit in der Rechtsprechung permanent ergänzt und verfeinert.

In Inhalt und Form entsprechen die im Auslegungsvorgang erarbeiteten definitorischen Elemente dem Typus der terminologischen Definition jedoch weit weniger als Legaldefinitionen. Während es sich bei Legaldefinitionen um weitgehend kontextfrei interpretierbare und monofunktionale Aussagen handelt, sind definitorische Elemente in Gerichtsurteilen Bestandteile kohärenter Texte. Zudem können (außer im Fall höchstrichterlicher Entscheidungen) im Rahmen der Auslegung getroffene Definitionen von übergeordneten Gerichten überprüft und akzeptiert oder abgelehnt werden. Wenn ein Richter also weitreichende begriffliche Festlegungen trifft, läuft er Gefahr, dass seine Entscheidung im Nachhinein "kassiert" wird, weil diese sich als nicht haltbar erweisen. Allein schon aus diesem Grund werden richterliche Definitionen oft möglichst vorsichtig formuliert und in ihrer Reichweite stark eingeschränkt. Dennoch entwickeln viele richterliche Bedeutungsfestlegungen und Präzisierungen eine hohe faktische Bindungswirkung in der späteren Rechtsprechung.

Wir werden im Folgenden am Beispiel von zwei Auszügen aus Gerichtsurteilen illustrieren, in wie hohem Maße in der Rechtsprechung Legaldefinitionen durch eine "zweite Schicht" definitorischer Information ergänzt werden, die permanent wieder aufgegriffen und fortgeschrieben wird. Wir möchten dabei

[16] Zur Beschreibung des Zusammenhangs zwischen begrifflicher Offenheit und Flexibilität des Rechtssystems spricht Hart (1961), 123 ff., in Analogie zu dem in Waismann (1945) diskutierten *open texture* natürlich-sprachlicher Begriffe von einem *open texture* der Normen der Rechtsordnung. Zu Gemeinsamkeiten und Unterschieden zwischen den von Waismann und Hart verwendeten Begrifflichkeiten vgl. Bix (1995), 7-36

auch das bisher über die Bestimmung von Begriffen in Urteilstexten Gesagte anhand zusammenhängender Textpassagen nachvollziehen. Außerdem werden wir schon – im Vorgriff auf eine detailliertere Untersuchung im nächsten Kapitel – exemplarisch einige der sprachlichen Mittel aufzeigen, mit denen präzisierende Ausdifferenzierungen und partielle Neufestlegungen vorgenommen werden, mit denen also neues begriffliches Wissen erzeugt wird.

Die zitierten Stellen sind so ausgewählt, dass sie zwei typische "Begründungssituationen" repräsentieren, in denen intensiv semantisch argumentiert wird: Im ersten Beispiel ist die Einordnung des Sachverhalts umstritten, kann jedoch durch Präzisierungen auf der Grundlage der bisherigen Rechtsprechung gerechtfertigt werden. Im zweiten Beispiel ist zur Behandlung des Falls eine echte Neufestlegung durch das Gericht erforderlich, die durch ein teleologisches Argument gestützt wird.

### 1.3.1 Umstrittener Fall

Wir betrachten nun einen Ausschnitt der Begründung einer Entscheidung des 3. Senats des BFH vom 10.6.1988 (III R 65/84). Es war festzustellen, ob ein auf einer Betonplatte aufgestellter Bürocontainer bewertungsrechtlich als Gebäude anzusehen ist. Die Betreiberin eines Großhandels hatte Klage gegen das Finanzamt erhoben, das ihr eine Investitionszulage gemäß § 19 des Berlinförderungsgesetzes für die Anschaffung eines Bürocontainers mit der Begründung versagt hatte, es handle sich bewertungsrechtlich nicht um ein bewegliches Wirtschaftsgut, sondern um ein Gebäude. Die Klägerin widersprach unter anderem mit dem Argument, der Container verfüge über kein Fundament. Die Betonplatte, auf der er aufgestellt sei, stelle kein Fundament dar, denn sie sei nicht frostsicher im Boden verankert, sondern nur auf den Untergrund gegossen.

Das FG Berlin hatte der Klage stattgegeben, das Finanzamt hatte dann gegen diese Entscheidung Revision eingelegt. Die für die Revision relevanten Argumente des FG Berlin fasst der BFH im Tatbestand seines Urteils folgendermaßen zusammen:

> Das Finanzgericht (FG) gab hingegen der Klage statt. Es ging davon aus, daß die Klägerin ein bewegliches Wirtschaftsgut angeschafft habe und daß der Container diese Eigenschaft auch nicht durch die Aufstellung auf dem Betriebsgelände der Klägerin verloren habe. Denn es fehle an einer festen Verbindung des Containers mit dem Grund und Boden. Auch sei kein Fundament vorhanden, auf dem der Container ruhe oder mit dem er verbunden sei.

Für diese Begründung war das FG Berlin zusätzlich davon ausgegangen, dass die Betonplatte unter dem Bürocontainer auch deshalb nicht als Fundament zu werten sei, weil sie hauptsächlich dem Ausgleich der Bodenschräge diene.

Der BFH wendet sich mit folgender Argumentation gegen die Ansicht des FG Berlin und der Klägerin aus der ersten Instanz (die Satznummern haben wir hinzugefügt):

> (1) Bewertungsrechtlich ist ein Bauwerk als Gebäude anzusehen, wenn es – neben anderen im Streitfall nicht zweifelhaften Merkmalen – fest mit dem Grund und Boden verbunden ist (vgl. BFH-Urteil vom 25. März 1977 III R 5/75, BFHE 122, 150, BStBl II 1977, 594).
>
> (2) Das ist der Fall, wenn einzelne oder durchgehende Fundamente vorhanden sind, das Bauwerk auf diese gegründet und dadurch mit dem Boden verankert ist (s. z.B. das BFH-Urteil vom 21. Februar 1973 II R 140/67, BFHE 109, 156, BStBl II 1973, 507; vgl. auch das BFH-Urteil vom 4. Oktober 1978 II R 15/77, BFHE 126, 481, BStBl II 1979, 190).
>
> (3) Befindet sich das Bauwerk auf einem Fundament, so ist es unerheblich, ob es mit diesem fest verbunden ist (vgl. BFH-Urteil vom 3. März 1954 II 44/53 U BFHE 58, 575, BStBl III 1954, 130).
>
> (4) Ausreichend ist es u.U. auch, daß ein Bauwerk kraft seiner Eigenschwere auf den Trägerelementen ruht (vgl. BFH-Urteil vom 18. Juni 1986 II R 222/83, BFHE 147, 262, BStBl II 1986, 787, mit zahlreichen Hinweisen).
>
> [. . . ]
>
> (5) Der Begriff des Fundaments ist im Gesetz und auch in Abschn.7 Abs.1 Satz 1 der Abgrenzungs-Richtlinien vom 31. März 1967 (BStBl II 1967, 127) nicht definiert.
>
> (6) Nach der Rechtsprechung ist dieser Begriff nicht eng auszulegen;
>
> (7) es genügt für die Annahme eines Fundaments vielmehr jede gesonderte (eigene) Einrichtung, die eine feste Verbindung des aufstehenden “Bauwerks” mit dem Grund und Boden bewirkt (vgl. Urteil in BFHE 100, 570, BStBl II 1971, 161, mit weiteren Rechtsprechungshinweisen).
>
> (8) Auf die Tiefe des Fundaments kommt es nicht an; auch nicht auf das Material, aus dem das Fundament besteht (vgl. Senatsurteil

> vom 19. Januar 1962 III 228/59 U, BFHE 74, 315, BStBl III 1962, 121).
>
> (9) Eine feste Verbindung mit dem Grund und Boden wird durch einen gegossenen Zementboden auch dann hergestellt, wenn dieser sich nicht in der Erde oder im Boden befindet (vgl. Senatsurteil vom 24. Mai 1963 III 140/60 U, BFHE 77, 156, BStBl III 1963, 376, 377, m.w.N.).
>
> [. . . ]
>
> (10) Daß der Betonuntergrund auch zum Ausgleich der Schräglage des Erdbodens dient, hindert – entgegen der Beurteilung durch das FG – die Annahme eines Fundaments nicht.
>
> (11) Aufgabe eines Fundaments ist es, die aus dem Bauwerk herrührenden Lasten und Kräfte so in den Baugrund abzuleiten, daß für das Bauwerk die Standsicherheit gewährleistet ist (vgl. Meyers, Enzyklopädisches Lexikon, 9.Aufl., Stichwort: Grundbau, sowie Meyers Großes Universallexikon, Stichwort: Fundament).

Bei der vorzunehmenden Einordnung (*Container* als *Gebäude*) handelt es sich um einen "schwierigen Fall". Wie die zitierte Textpassage erkennen läßt, hat in dieser Situation die Auseinandersetzung mit Gegenargumenten, insbesondere mit den vorgegebenen Argumentationslinien der Konfliktparteien und der Vorinstanz, Vorrang vor dem Aufbau einer geradlinigen Hypothesenkette.[17] Es wird intensiv auf Gründe *und* Gegengründe zum Vorliegen der einzelnen Tatbestandsmerkmale eingegangen.

Dabei werden gezielt Besonderheiten des Falls untersucht, die im Hinblick auf die betrachteten Fallklassen problematisch sind. Diese bilden Ansatzpunkte zur vergleichenden Beurteilung des Sachverhalts vor dem Hintergrund ähnlicher Fälle. Zugleich identifizieren die getroffenen Festlegungen wiederum geeignete Eigenschaften für die Analogiebildung in neuen Fällen.[18] Sie enthalten daher – auch wenn kein neues Konzept eingeführt wird – über den Einzelfall hinaus relevante begriffliche Information.

[17] Zu entnehmen ist diese Regel u.a. Lehrbüchern, die gezielte praktische und stilistische Hinweise zur Abfassung von Urteilen geben. Beispielsweise schreibt Sattelmacher und Sirp (1989): *Innerhalb der Grenzen, die sich aus der Beschränkung der Urteilsbegründungen auf das allein für die Entscheidung erhebliche ergeben, sollen die Entscheidungsgründe den Sach- und Streitstoff erschöpfen. Der Richter muss das Vorbringen der Parteien nicht nur zur Kenntnis nehmen, sondern auch in den Entscheidungsgründen verarbeiten. Dies erfordert der Anspruch der Parteien auf rechtliches Gehör* (261, Hervorhebungen im Original)

[18] Vgl. zum *typisierenden Fallvergleich* in der Auslegung Zippelius (2005), 72 ff.

Schon den Ausgangspunkt der Argumentation bildet im Beispiel keine Legaldefinition, sondern eine in der Rechtsprechung (schon des Reichsfinanzgerichtshofs) etablierte Begriffsbestimmung:

> Nach der ständigen Rechtsprechung des RFH ist ein Gebäude ein Bauwerk, das durch räumliche Umfriedung Personen und Sachen Schutz gegen äußere Einflüsse gewährt, den Eintritt von Menschen gestattet und von einiger Beständigkeit ist [...]. Das Bauwerk muß mit dem Grund und Boden fest verbunden sein.
>
> (RFH, 15. Mai 1941, III 43/41)

Dass der BFH im Weiteren konsequent an Präzisierungen aus anderen vergleichbaren Fällen anknüpft, zeigen die in fast jedem Satz enthaltenen Verweise auf frühere Entscheidungen. Auch die in dem angeführten Urteil herausgearbeiteten Festlegungen wurden fortgeschrieben. So begründete schon wenige Monate später der BFH bei einem auf vier einzeln in den Boden eingelassenen Betonklötzen ruhenden Container die Annahme eines Fundaments ebenfalls mit der Kraftableitungsfunktion der Stützkonstruktion. Er berief sich dabei auf die von uns zitierte Entscheidung (BFH 3.Senat vom 23.09.1988, III R 9/85). In der weiteren Rechtsprechung wurden die Feststellungen zu den hier angesprochenen Aspekten für verschiedene Leichtbaukonstruktionen weiter präzisiert (z.B. muss es sich bei den abzuführenden Lasten nicht um statische Lasten handeln, es kommen auch Zug- und Sogwirkungen durch Windkräfte in Frage, vgl. BFH 2. Senat vom 20.09.2000, II R 60/98; dagegen darf ein Fundament nicht durch bloßen Abtransport zu entfernen sein, vgl. BFH 3. Senat vom 23.09.1988, III R 67/85).

Desweiteren ist an der zitierten Passage vor allem interessant, wie einzelne Merkmale fokussiert und ausdifferenziert werden. Schon bei der Anführung der *Gebäude*-Definition in Satz 1 greift der BFH nur ein einzelnes Merkmal (*feste Verbindung mit Grund und Boden*) heraus und begründet diese Selektion mit der Charakteristik des Falls: Die weiteren Merkmale werden als *im Streitfall nicht zweifelhaft* ausgeblendet. Satz 2 greift die für den vorliegenden Fall relevante Fallklasse heraus (*Gebäude mit Fundament*). So wird weiter eingeschränkt, welche Aspekte im Folgenden näher zu untersuchen sind. Die anschließende Argumentation geht von der Definition des Begriffs *Gebäude* auf den Begriff *Fundament* über und fokussiert hier wiederum die für den Fall relevante Unterklasse *Fundament aus gegossenem Zement*.

Hinreichende und notwendige Bedingungen ergänzen sich für keines der Konzepte zu einer vollständigen Definition. Viele der getroffenen Feststellungen liefern zudem weder hinreichende noch notwendige Bedingungen, sondern

lehnen in Erwägung gezogene Bedingungen ab – es werden dialektisch Gegenargumente abgehandelt. Effektiv werden so die für den Fall relevanten Randbereiche der untersuchten Begriffe beleuchtet, und die Argumentation wird von der Definition des RFH auf den konkreten Fall hingeführt. Dabei wird jedoch der Anspruch der Allgemeingültigkeit auch in den stark fallbezogenen Äußerungen aufrecht erhalten.

Als sprachliche Besonderheiten der zitierten Passage fallen auf:

- die häufige Verwendung indefiniter Nominalphrasen (z.B. *ein Bauwerk*, *eine feste Verbindung*) bzw. Nominalphrasen mit dem definiten Artikel, die diese wieder aufgreifen (etwa *das Bauwerk* in Satz 2). Mit diesen Mitteln wird die Allgemeingültigkeit der Begriffsbestimmung zum Ausdruck gebracht und der Zusammenhang zwischen den angeführten Bedingungen etabliert.

- die Häufung von Partikeln und Adverbien (z.B. *auch*, *dann*, *vielmehr*), mit denen die Geltung, die logische Funktion und die Beziehung der einzelnen Bedingungen untereinander ausdifferenziert werden.

- Als *definitorisch* markiert ist ein Großteil der Sätze durch die verwendeten Prädikate. Diese tragen meistens gleichzeitig (wie *ansehen als*, *unerheblich sein*, *genügen für*) auch zur Ausdifferenzierung des logischen Status der ausgedrückten Anwendungsbedingungen bei.

### 1.3.2 Neufestlegung in einem abweichenden Fall

In dem eben betrachteten Fall folgte die vorgenommene Klassifikation zwar nicht allein aus Legaldefinitionen, war aber durch eine Kombination *präzisierender Feststellungen* zu rechtfertigen, die sich weitestgehend auf die bisherige Rechtsprechung stützen konnten. Es kommt jedoch auch vor, dass weder Legaldefinitionen noch der in der Rechtsprechung zu einer Norm etablierte "definitorische Bestand" zu einer Entscheidung führen. Diese Situation kann insbesondere dann entstehen, wenn Fälle auftauchen, die insgesamt nicht vorauszusehen waren, als eine Norm (mit den enthaltenen Begriffsbestimmungen) aufgestellt wurde. In solchen Fällen ist eine echte *Bedeutungsfestlegung* zu treffen und zu rechtfertigen. Die Rechtfertigung kann dann etwa durch systematische Erwägungen oder in einer teleologischen Argumentation aus dem Normzweck erfolgen.

Dies war beispielsweise der Fall in einer Entscheidung des 5. Senats des BFH (18.8.2005, V R 50/04). "Disk Jockeys" spielten bei eine Veranstaltung mittels spezieller CD-Player und Schallplattenspieler Musik von Tonträgern ein. Dabei

wurde die Musik aber auf verschiedenste Art gemischt und stark individuell variiert. Festzustellen war nun, ob eine solche Veranstaltung noch als Konzert im Sinne des Umsatzsteuergesetzes zu begünstigen ist. Das zuständige Finanzamt hatte Revision gegen eine Entscheidung des FG Berlin eingelegt, nach der die Veranstaltung als Konzert anzuerkennen war.

Ausgangspunkt der Entscheidungen sowohl des BFH als auch der Vorinstanz ist die im Umsatzsteuerrecht gängige Rechtsprechungsdefinition des Konzertbegriffs, die der BFH folgendermaßen formuliert:

> Konzerte i.S. des § 12 Abs. 2 Nr. 7 Buchst. a UStG 1993 sind Aufführungen von Musikstücken, bei denen Instrumente und / oder die menschliche Stimme eingesetzt werden.

Das Finanzamt hatte nun argumentiert, bei den im Streitfall verwendeten CD-Playern und Plattenspielern handle es sich nicht um Instrumente, weshalb die Veranstaltung nicht als Konzert im einschlägigen Sinne anzusehen sei. Zur Begründung dieser Ansicht hatte es eine am alltagssprachlichen Verständnis orientierte Definition des Ausdrucks *Musikinstrument* postuliert:

> Musikinstrumente seien nur solche Gegenstände, die mit dem Ziel konstruiert oder verändert würden, mit ihnen Musik erzeugen zu können. Weder Abspielgeräte für Tonträger noch Mischpulte oder ähnliche technische Geräte seien Musikinstrumente, weil diese ohne einen Tonträger keine Töne erzeugen könnten.

Das vom Finanzamt vorgebrachte Argument ist offenkundig nicht schlüssig. Die vorgeschlagene Definition des Begriffs *Musikinstrument* schließt nicht aus, dass zur Erzeugung von Tönen neben dem Instrument selber noch weiteres Zubehör benötigt wird. Allerdings ergibt die Definition auch nicht zwingend, dass CD-Player und Plattenspieler unter den Begriff fallen.

Der BFH begegnet der Argumentation des Finanzamts mit folgenden Ausführungen:

> (1) Der Senat hält an dieser Auslegung des Begriffs “Konzert” fest; danach ist das bloße Abspielen eines Tonträgers kein Konzert.
>
> (2) Jedoch bedarf der Begriff “Instrument” angesichts der technischen Entwicklungen auf dem Gebiet der Musik einer Präzisierung.
>
> (3) Um den Wettbewerb zwischen neuen und bestehenden Musiktechniken sowie Musikrichtungen umsatzsteuerrechtlich nicht zu

> behindern, können bei Musik, die – wie es das FG zu den Musikrichtungen "Techno" und "House" festgestellt hat – in wesentlichen Teilen durch Verfremden und Mischen bestehender Musik komponiert wird, Plattenteller, Mischpulte und CD-Player "Instrumente" sein, mit denen die Musik im Rahmen eines Konzerts dargeboten wird, wenn sie (wie konventionelle Instrumente) zum Vortrag des Musikstücks und nicht nur zum Abspielen eines Tonträgers genutzt werden.
>
> (4) Dass es sich dabei nicht um Musikinstrumente im üblichen Sinne handelt, ist unerheblich.

Die Problematik des Falls liegt – wie vom BFH mit Satz 2 und Satz 4 der zitierten Passage festgestellt – darin, dass der Begriff *Musikinstrument* im üblichen Verständnis *angesichts der technischen Entwicklungen im Bereich der Musik* zu undifferenziert ist, um eine klare Entscheidung zu ermöglichen. Aus diesem Grund ist auch die vom Finanzamt vorgeschlagene Definition für die Falllösung unergiebig.

Satz 3 präzisiert den Begriff daher durch eine ausweitende Modifikation der vom Finanzamt ins Spiel gebrachten Definition. Begründet wird dies mit einem zugleich teleologischen und systematischen Argument: Nur durch die vorgenommene Präzisierung kann nach der Ansicht des Gerichts die steuerliche Benachteiligung von Musikrichtungen vermieden werden, die sich neue Erzeugungstechniken zu Nutze machen, ohne von der für den Konzertbegriff in der Rechtsprechung etablierten Definition abrücken zu müssen.

Der BFH trifft hier eine eigene Festlegung, die er nicht durch Belege aus der früheren Rechtsprechung stützt. Unter welchen Bedingungen solche Festlegungen methodologisch zulässig sind und ab welchem Punkt sie Verletzungen der Wortlautgrenze darstellen, kann hier nicht generell diskutiert werden.[19] Für den vorliegenden Fall dürfte es wichtig sein, dass die getroffene Festlegung nicht im Widerspruch zu einer Legaldefinition, sondern nur zu einer (vermuteten) alltagssprachlichen Verwendungsregel steht. Zudem ist die aus der Erweiterung des *Instrument*-Begriffs resultierende Klassifikation der beschriebenen Techno-Veranstaltung als Konzert nicht nur erwünscht, sondern auch ohne weiteres mit der alltagssprachlichen, von der Rechtsprechungsdefinition unabhängigen Verwendungsweise dieses Begriffs vereinbar.

Da der BFH seine Festlegung eigenverantwortlich und ohne Rückgriff auf andere Quellen, vornimmt, schwächt er den mit der Präzisierung verbundenen Geltungsanspruch stark ab. In sprachlicher Hinsicht äußert sich dies darin, dass der Anwendungsbereich der präzisierenden Aussage durch eine zwar generisch

[19] Vgl. hierzu Zippelius (2005), 83 ff.

ausgedrückte, aber inhaltlich ausgesprochen spezifische Beschreibung der Situation im vorliegenden Fall sehr eng eingegrenzt wird (*Musik wird analog zu den Methoden bei "Techno" und "House"-Musik in wesentlichen Teilen durch Verfremden und Mischen bestehender Musik komponiert*). Die modal abgetönte Formulierung *können… genutzt werden* schränkt die Verbindlichkeit der Festlegung noch zusätzlich ein. Trotz dieser Einschränkungen weist die vorgenommene Präzisierung jedoch über den konkreten Fall hinaus. Dies zeigt sich auch daran, dass auch dieses Urteil bereits kurze Zeit später in anderen Fällen zur Stützung der Argumentation herangezogen wurde (z.B. zur Abgrenzung bei der Bewertung einer reinen Disco-Veranstaltung – BFH 5. Senat, 12.01.2006, V R 67/03 – oder als analoger Fall in einem Urteil zur umsatzsteuerlichen Bewertung neuartiger Behandlungsmethoden eines heilkundlichen Psychotherapeuten als Heilbehandlung – FG Köln 10. Senat, 19.01.2006, 10 K 5354/02).

### 1.3.3 Fazit

Wir haben in diesem Kapitel zunächst einen kurzen Überblick über verschiedene Aspekte der juristischen Begründungslehre und der Regeln des Auslegungsprozesses gegeben, die für unsere Arbeit relevant sind. Vor diesem Hintergrund haben wird dann vor allem die tastende, argumentativ gesteuerte Ausarbeitung der Semantik fallrelevanter Begriffe in Entscheidungsbegründungen näher betrachtet. Schließlich haben wir zwei längerere definitorische Textpassagen (Auszüge aus zwei Entscheidungen des BFH) untersucht.

Diese exemplarische Untersuchung hat konkret gezeigt, wie in Urteilsbegründungen ein vorgegebener Bestand an Definitionen um eine "zweite Schicht" weiterer Bedeutungsfeststellungen ergänzt wird. Im Gegensatz zu Legaldefinitionen handelt es sich bei solchen Angaben oft um stark eingeschränkte und auf den ersten Blick nur im gegebenen Kontext relevante Aussagen. Jedoch sind richterliche Bedeutungsfeststellungen, auch wenn sie der Formulierung nach stark eingeschränkt und kontextgebunden gefasst sind, keineswegs nur in ihrem ursprünglichen Kontext und für einen konkreten Fall relevant.

Sie entwickeln vielmehr – obwohl formal nur Legaldefinitionen volle normative Bindungswirkung zukommt – häufig über den Fall hinaus große Verbindlichkeit. Die betrachteten Beispiele haben gezeigt, in welch hohem Maße bei der Auslegung Ergebnisse der bisherigen Rechtsprechung zitiert und verwertet werden. Teilweise geschieht dies wohl aus textökonomischen Erwägungen, etwa wenn für ausführliche Darstellungen auf andere Urteile verwiesen wird, anstatt ganze Passagen zu wiederholen. Im Falle definitorischer Feststellungen besteht eine besonders wichtige Funktion von Rechtsprechungszitaten jedoch darin, Berührungspunkte aufzuzeigen, die das Gericht zwischen dem vorliegenden Fall und Sachverhalten aus der früheren Rechtsprechung sieht. Analogieba-

siert werden so für bestimmte Fallklassen relevante Merkmale identifiziert und Verallgemeinerungen gewonnen, die es erlauben, den konkreten Fall auch dann nach einem generalisierbaren Prinzip zu entscheiden, wenn ein solches dem Gesetz nicht direkt zu entnehmen ist. Gleichzeitig kann durch Bezugnahme auf entsprechende Quellen die Autorität der Argumentation des Gerichts gestärkt werden.

Bei Urteilsbegründungen handelt es sich um hochgradig geplante, strukturierte und zielgerichtete argumentative Texte. Die zitierten Passagen lassen deutlich erkennen, dass richterliche Definitionen als Bestandteile dieser Texte mit entsprechend differenzierten sprachlichen Mitteln zum Ausdruck gebracht werden. Im nächsten Kapitel untersuchen wir den Zusammenhang zwischen Form und Funktion solcher Feststellungen aus linguistischer Sicht. Wir verschaffen uns dabei einen umfassenderen Überblick über die Bandbreite der verwendeten Ausdrucksformen.

# Kapitel 2

# Sprachliche Realisierung von Definitionen

Der Definitionsbegriff in seiner vollen Bandbreite entzieht sich selbst einer klaren Definition. Wie in 1.2.1 ausführlicher besprochen, wird von Definitionen traditionell meist gefordert, dass sie das *Äquivalenzkriterium* und das *Essentialitätskriterium* erfüllen. Während das Äquivalenzkriterium die Vollständigkeit und Trennschärfe einer Definition absichert, stellt das Essentialitätskriterium sicher, dass die angegebene Information für das Definiendum tatsächlich zentral ist.

Gestützt auf die Sicht pragmatisch orientierter sprachtheoretischer Ansätze lässt sich gegen eine solche enge Eingrenzung des Definitionsbegriffs einwenden, dass jede Verwendungsinstanz eines Begriffs unter Umständen zugleich als ein Beispiel für seine Verwendung dienen kann. Unter einer solchen Annahme wäre also jeder Satz potentiell definitorisch, da er Auswirkungen auf die zukünftige Verwendung der benutzten Begriffe haben kann (diese Position vertritt im juristischen Bereich in extremer Form die in 1.1.3 diskutierte Strukturierende Rechtslehre).

Als Grundlage für den Aufbau eines definitionsbasierten juristischen Informationssystems ist jedoch weder der durch Äquivalenz- und Essentialitätskriterium charakterisierte Definitionsbegriff noch die angesprochene radikalpragmatische Alternative geeignet. Ein Benutzer, der die Definition eines Begriffs in einer Rechtsdatenbank nachschlägt, ist in der Regel nicht an einer Auflistung aller Sätze interessiert, die den fraglichen Begriff enthalten. Andererseits muss die in diesem Fall relevante Information nicht immer und in allen denkbaren Zusammenhängen logisch äquivalent zum Suchbegriff sein. Eine Textpassage ist als Suchergebnis von Interesse, wenn sie vom Autor, also dem Gericht, als Feststellung zur Semantik eines Begriffs intendiert ist und außerdem prinzipiell eine Übertragung in neue Verwendungszusammenhänge ermöglicht. Für die Umsetzung in einem juristischen Informationssystem müssen diese Anforderungen zum einen weiter präzisiert werden. Zum anderen ist entscheidend, ob sich sprachliche Indikatoren identifizieren lassen, die eine automatische Suche nach definitorischen Textpassagen im Sinne des so gewonnenen “erweiterten Definitionsbegriffs” ermöglichen.

In diesem Kapitel versuchen wir zuerst auf konzeptueller Ebene einen für den juristischen Zusammenhang sinnvollen weit gefassten Definitionsbegriff zu charakterisieren. Zu diesem Zweck analysieren wir in 2.2, wie und aus welchen Gründen definitorische Aussagen in Urteilsbegründungen in inhaltlich-funktionaler Hinsicht von dem durch das Essentialitäts- und Äquivalenzkriterium charakterisierten Idealtypus abweichen.

In 2.3 untersuchen und systematisieren wir dann die Bandbreite der Formulierungsmuster, die in der Rechtssprache für definitorische Aussagen verwendet werden. In 2.4 schließlich diskutieren wir die Ergebnisse einer Studie, bei der ein Korpus aus sechzig Entscheidungstexten unter kontrollierten Bedingungen doppelt auf definitorische Passagen annotiert wurde. Die so erzeugte Ressource dient uns in den folgenden Kapiteln als Goldstandard für die Evaluation unseres Systems zur automatischen Definitionsextraktion. Zunächst gehen wir jedoch (in 2.1) kurz auf die Datengrundlage ein, auf die sich unsere Erörterungen in 2.2 und 2.3 stützen (eine Aufstellung aller im Rahmen unserer Arbeit verwendeten Korpora findet sich in Anhang A).

## 2.1 Annotation: Datengrundlage und Methode

### 2.1.1 Korpusaufbau

Die in diesem Abschnitt präsentierten Ergebnisse beruhen auf der explorativen Untersuchung eines Korpus aus vierzig Gerichtsentscheidungen aus dem Datenbestand der Firma *juris* zu verschiedenen (hauptsächlich dem Verwaltungsrecht zuzuordnenden) Sachgebieten.[1] Die Verteilung der untersuchten Entscheidungen auf Gerichtsbarkeiten sowie deren genauere Zuordnung zu übergeordneten Sachgebietskategorien[2] gibt Tabelle 2.1 wieder.

In der Regel finden sich nur in den begründungsbezogenen Dokumentteilen von Urteilstexten, also vor allem in den *Urteilsgründen* und *Leitsätzen*, juristisch relevante Definitionen (für Näheres zum Aufbau von Entscheidungsdokumenten vgl. 4.2.1). In den als *Tatbestand* betitelten Abschnitten mit der Beschreibung des jeweils zu entscheidenden Sachverhalts treten dagegen kaum Definitionen auf. Außerdem geben juristische Argumentationen in diesen Abschnitten generell nur die Meinung einer der Streitparteien wieder, die unter

[1] Wir möchten der Firma *juris* an dieser Stelle dafür danken, dass sie uns für unsere Forschungen den Zugang zu ihrem Datenbestand ermöglicht hat.

[2] Bei den angegebenen Sachgebieten handelt es sich um die obersten Kategorien, denen die betrachteten Entscheidungen in der von *juris* verwendeten Systematik zugeordnet sind. Ein Dokument kann dabei mehreren Kategorien angehören, so dass sich in der rechten Spalte von Tabelle 2.1 eine Summe von mehr als vierzig ergibt.

| Gerichtsbarkeit | Anzahl | Sachgebiet | Anzahl |
|---|---|---|---|
| BVerfG | 2 | Arbeitsrecht | 1 |
| BVerwG | 3 | Allgemeines Verwaltungsrecht | 7 |
| Landesverwaltungsgerichte | 14 | Bürgerliches Recht | 9 |
| Verwaltungsgerichte | 3 | Besonderes Verwaltungsrecht | 35 |
| OLG | 7 | Europarecht | 2 |
| LG | 5 | Weitere Gerichtsverfahrensordnungen | 6 |
| Landesarbeitsgerichte | 2 | Gerichtsverfassung und -zuständigkeit | 2 |
| Landessozialgerichte | 1 | Handels- und Wirtschaftsrecht | 14 |
| Ausländische oberste Gerichte | 3 | Regionen, Rechtswissenschaft, Kirchenrecht | 23 |
| | | Sozialrecht | 1 |
| | | Strafrecht | 1 |
| | | Staats- und Verfassungsrecht | 10 |
| | | Zivilprozessrecht | 7 |
| | | Recht des öffentlichen Dienstes | 3 |

Tabelle 2.1: Zuordnung der Entscheidungen aus unserer explorativen Studie (*Pilotstudien-Korpus*) zu Gerichtsbarkeiten und Sachgebietskategorien

Umständen in der endgültigen Entscheidung teilweise oder vollständig abgelehnt wird.

### 2.1.2 Annotation

Für unsere Untersuchung haben wir daher aus den vierzig zu Grunde gelegten Entscheidungen vorab nur die Entscheidungsbegründungen und Leitsätze extrahiert. Dieses Korpus (im Folgenden auch: *Pilotstudien-Korpus*) legen wir hier in einer aufbereiteten Form zugrunde. Zunächst erfolgte eine automatische Textanalyse mittels eines Satz- und Wortgrenzenerkenners (vgl. hierzu 4.2). In dem so vorverarbeiteten Textbestand (127 349 Wörter in 3757 Sätzen) wurden:

1. definitorische Passagen annotiert und durch den Annotator je nach Konfidenz in seine Annotationsentscheidung als *sicher*, *unsicher* oder *zweifelhaft* markiert

2. Sätze in solchen Passagen den “Informationsschichten” *Kerndefinition* bzw. *elaborierende Information* zugeordnet (vgl. die einleitenden Erläuterungen in 2.2)

3. die argumentative Funktion der einzelnen Sätze bestimmt (vgl. 2.2.1)

Die Annotation erfolgte für die Zwecke der hier zunächst diskutierten explorativen Studie einfach. Für die definitorischen Passagen sowie die Zuordnung der einzelnen Sätze zu "Informationsschichten" wurde sie durch einen Juristen (nach dem ersten Staatsexamen) vorgenommen. Die für die Bestimmung der argumentativen Satzfunktionen verwendeten Kategorien wurden erst im Rahmen der Auswertung der ersten beiden Annotationsschichten erarbeitet und konsolidiert. Die entsprechende Information wurde nachträglich durch den Autor der Arbeit annotiert.

Die gewonnenen Erfahrungen wurden dann in Form eines Annotationsleitfadens zusammengefasst, auf dessen Grundlage – wie in 2.4 beschrieben – ein umfangreicheres Korpus doppelt annotiert wurde. Dabei wurde auch die Annotatorenübereinstimmung (sog. *Inter-Annotator-Agreement*, IAA) ermittelt.

Im Pilotstudien-Korpus identifizierte der Annotator insgesamt 138 definitorische Textsegmente. Er stufte seine Entscheidung in 87 Fällen als *sicher* ein, 35 Textspannen erhielten den mittleren Konfidenzwert (*unsicher*), die verbleibenden entfielen auf die Klasse *zweifelhaft*.

Die kleinsten annotierbaren Einheiten waren Einzelsätze. Dies erwies sich nicht als problematisch: Von den 176 als definitorisch identifizierten Sätzen enthalten nur elf mehrere Definitionen. Für die im Folgenden beschriebenen Auswertungen haben wir in diesen Fällen nur die jeweils prominenteste Definition im Satz in Betracht gezogen. Dem Textaufbau nach zusammengehörige Sätze mit definitorischer Information zu demselben Begriff wurden in der Annotation zu einem Definitionskomplex zusammengefasst. In zehn Fällen betraf dies Sätze, die im Text nicht direkt aufeinander folgten. Durchschnittlich umfassten die annotierten Definitionen 1,4 Sätze.

Weder unser Pilotstudien-Korpus noch das doppelt annotierte Goldstandard-Korpus sind umfangreich genug, um Schlussfolgerungen zu Besonderheiten rechtssprachlicher Definitionen statistisch im wünschenswerten Maße absichern zu können. In jedem Fall erlauben sie aber Aussagen über über Tendenzen, zu deren praktischer Validierung wir nicht zuletzt mit der Entwicklung und Evaluation unseres Definitionsextraktionsverfahrens in Kap. 5 beitragen.

In der linguistischen Literatur fehlen systematische statistische Untersuchungen zu Definitionen bisher völlig. Die Möglichkeit eines quantitativen Vergleichs unserer Ergebnisse mit anderen Studien (etwa zu Definitionen wissenchaftlich-technischen Texten) scheidet deshalb aus. Qualitativ zeigt sich aufgrund der Beobachtungen, die wir im Rahmen unserer Annotation gemacht haben, jedoch deutlich, dass sowohl die Vorgehensweise als auch die Formulierungsmuster bei der Begriffsbestimmung in Urteilsbegründungen oft deutlich von den in der Wissenschaftstheorie sowie der Terminologiewissenschaft und Lexikographie diskutierten Schemata "kunstgerechten" Definierens abweichen. Im Folgenden gehen wir zunächst auf die wichtigsten dieser Abweichungen in

inhaltlich-funktionaler Hinsicht ein und versuchen dann, auf der Grundlage der in unserer Studie vorgefundenen definitorischen Formulierungen verschiedene sprachliche Realisierungstypen juristischer Begriffsbestimmungen zu identifizieren.

## 2.2 Inhaltlich-funktionale Klassifikation rechtssprachlicher Definitionen

In inhaltlich-funktionaler Hinsicht besteht der Hauptunterschied zwischen terminologischen Definitionen in Wissenschaft und Technik und Definitionen in Urteilsbegründungen darin, dass Begriffe in Urteilsbegründungen in aller Regel nicht vor ihrer Verwendung kontextfrei und abstrakt definiert werden können. Sie müssen vielmehr im Zusammenhang eines kohärenten Textes bestimmt werden (siehe hierzu 1.2.1).

**Informationsschichten.** Die in juristischen Argumentationen zu diesem Zweck verwendete definitorische Information kann generell zwei (bereits eingangs erwähnten) *Schichten* zugeordnet werden:

1. Definitorische Kernangaben, die darauf abzielen, das zu definierende Konzept zugleich konzise zu fassen und in all seinen wesentlichen Aspekten abzudecken.

2. Elaborierende Angaben, die einzelne Bedeutungsaspekte vertiefen und genauer bestimmen.

Die aus einem in unserer explorativen Studie analysierten Urteil stammende Textpassage in (2.1) verdeutlicht diese Einteilung (die Sätze sind im Original nicht nummeriert):

(2.1) (1) Eine Rundfunksendung iSv UrhG AUT §§ 17, 59a liegt immer schon dann vor, wenn ein Werk mit Hilfe von Hertz'schen Wellen innerhalb der Reichweite dieser Wellen jedem wahrnehmbar gemacht wird, der sich eines entsprechenden Empfangsgeräts bedient.

(2) Gleichgültig ist, ob die Sendung auch wirklich wahrgenommen wird;

(3) es genügt, daß die Möglichkeit hierzu geboten wird.

(Oberster Gerichtshof Wien, 13. November 2001, AZ 4 Ob 182/01w, juris)

Der erste Satz dieses Abschnitts liefert hinreichende und notwendige Anwendungskriterien des Begriffs *Rundfunksendung*. Er bestimmt das Definiendum somit zwar umfassend und – logisch betrachtet – vollständig. Im konkreten Fall entscheidend ist aber der in dieser umfassenden Definition nicht thematisierte Aspekt *Wahrnehmung der Sendung*, den Satz (2) und (3) elaborieren.

Auch für die Klassifikation als Kern- bzw. elaborierende Angaben erwiesen sich in unserer Annotationsstudie Einzelsätze als geeignete Einheiten. Es ergab sich ein Verhältnis von etwa 2:1 zwischen den beiden Informationstypen (121 Sätze mit Kerninformation und 55 elaborierende Sätze bzw. 115 Kerndefinitionen und 53 elaborierende Aussagen bei Einschränkung auf die Konfidenzwerte *sicher* und *unsicher*). Alle 32 mehrsätzigen Definitionen umfassten sowohl Kerninformation als auch elaborierende Aussagen. Vielfach fanden sich Abfolgen, in denen so wie in (2.1) eine Kerndefinition durch elaborierende Angaben ergänzt wird. Allerdings bestanden auch 14 der insgesamt 138 identifizierten Definitionen nur aus elaborierenden Angaben.

Nur Kernangaben beinhalten Information, die generell (oder zumindest in einer großen Zahl von Fällen) selbständig für die Verwendung des definierten Begriffs maßgebend ist. Sie stützen sich meist auf Legaldefinitionen bzw. den durch gefestigte Rechtsprechung etablierten Bestand an terminologischem Wissen. Mit wissenschaftssprachlichen Definitionen teilen sie den relativ kontextfreien und kompakten Charakter. Oft erfüllen sie – wie auch in (2.1) – das Äquivalenzkriterium. Von den in unserer Annotationsstudie identifizierten Kerndefinitionen enthält mehr als die Hälfte hinreichende und notwendige Bedingungen. Bei einem Teil der verbleibenden Kerndefinitionen handelt es sich um unvollständig wiedergegebene, aber dem Juristen vollständig bekannte Definitionen (z.B. um konjunktive Bedingungen aus einer Legaldefinition, die aus Gründen der Inhaltsdisposition an unterschiedlichen Textstellen abgehandelt werden). Dies ist jedoch nicht immer so. Zum Teil sind Kerndefinitionen auch unvollständig, weil für den zu entscheidenden Fall im gegebenen Zusammenhang (z.B. für das Erreichen des Argumentationsziels) ein Konzept nicht vollständig definiert werden muss.

Elaborierende Angaben setzen die – explizit definierte oder in der vorgängigen Verwendung implizite – Semantik des Definiendum als in groben Zügen bekannt voraus. Vor diesem Hintergrund füllen sie eine verbleibende "Bestimmtheitslücke" (vgl. die Diskussion in 1.2.2) und liefern für die Begründung der Einzelfallentscheidung wichtige Bindeglieder. Sie stellen somit zwar keine allgemeine Referenz für die Verwendung des Definiendum dar, können aber in Fällen ausschlaggebend sein, die aufgrund der Kerndefinition nicht ohne weiteres zu entscheiden wären. Meistens sind sie stärker kontextuell eingebunden als Kerndefinitionen (in (2.1) beispielsweise durch elliptische Formulierung und anaphorische Elemente) und erfüllen typischerweise nicht das Äquivalenzkrite-

rium. In unserem Korpus handelt es sich bei sämtlichen elaborierenden Angaben um im logischen Sinne unvollständige Definitionen.

**Funktion.** Neben der Zuordnung zu den genannten Informationsschichten lassen sich bei vielen unvollständigen definitorischen Aussagen verschiedene Aspekte der *Funktion* unterscheiden, auf die wir im Folgenden anhand typischer Beispiele näher eingehen werden:

- Präzisierung einzelner Begriffsmerkmale
- Spezialisierte Bestimmung eines Begriffs durch fallbezogene Anwendungsbedingungen
- Kommentierende und erläuternde Zusatzangaben

Die Kategorien *Informationsschicht* und *Funktion* sind weder orthogonal zueinander noch vollkommen unabhängig: Während eine Präzisierung einzelner Begriffsmerkmale durch Kern- ebenso wie durch elaborierende Definitionen erfolgen kann, treten in den beiden anderen Funktionen typischerweise elaborierende Angaben auf. Auch handelt es sich bei den drei genannten funktionalen Gesichtspunkten nicht um untereinander ausschließende Kategorien. Ein erheblicher Teil der elaborierenden Angaben in unserem Korpus erfüllt gleichzeitig mehrere der angeführten Aufgaben.

### 2.2.1 Präzisierung einzelner Merkmale

Der erste funktionale Haupttypus von Definitionen präzisiert das Definiendum auf intensionaler Ebene. Dazu wird ein einzelnes Begriffsmerkmal fokussiert und detaillierter ausgearbeitet.

#### (a) Angabe von Grenzwerten

Dies kann zum einen dadurch geschehen, dass auf einer Skala ein Punkt gekennzeichnet wird, jenseits dessen sich ausschließlich klare (positive oder negative) Fälle hinsichtlich eines Merkmals befinden. Ist für die Anwendbarkeit des zu definierenden Begriffs eine messbare Größe entscheidend, wird zu diesem Zweck in der Regel wie in (2.2) numerisch ein Grenzwert fixiert. Auch bei Begriffen ohne metrische Komponente kann aber mit entsprechendem Hintergrundwissen oft ein sinnvolles Ordnungskriterium ausgemacht werden, hinsichtlich dessen ein “Grenzwert” angegeben werden kann. Dieser kann dann allerdings nicht numerisch festgelegt werden, sondern muss durch eine Beschreibung identifiziert werden, die u.U. selbst wiederum näher zu bestimmende (insbesondere oft *wertausfüllungsbedürftige*, vgl. 1.2.2) Begriffe enthält. So

werden in (2.3) zur Ausarbeitung des Merkmals *hässlich* Baumaßnahmen nach dem Grad der Einwirkung auf das ästhetische Empfinden des Betrachters geordnet, wobei eine Skala *beeinträchtigen* < *verletzen* angenommen wird.

(2.2) Bei einem Einfamilien-Reihenhaus liegt ein mangelhafter Schallschutz dann vor, wenn die Haustrennwand einschalig errichtet wurde und der Schalldämmwert von 57 dB nicht erreicht wird.

(OLG Stuttgart 1. Zivilsenat, 22. November 1995, AZ 1 U 199/93, juris)

(2.3) Eine Verunstaltung im Sinne dieser Regelung liegt nur vor, wenn ein häßlicher, das ästhetische Empfinden des Beschauers nicht nur beeinträchtigender, sondern verletzender Zustand geschaffen würde.

(Verwaltungsgerichtshof Baden-Württemberg 5. Senat, 12. August 1993, AZ 5 S 1018/92, juris)

#### (b) Präzisierung im argumentativen Zusammenhang

Außer durch Grenzwertangaben kann eine merkmalsbezogene Präzisierung auch durch Abwägungen zur Relevanz bestimmter abstrahierter Aspekte des Sachverhalts im Hinblick auf ein ausgewähltes Merkmal erfolgen. So knüpfen die Sätze 2 und 3 in (2.1) an die vorangehende Definition des Terminus *Rundfunksendung* an, indem sie das Merkmal *wahrnehmbar gemacht* fokussieren und die Bedeutung des Sachverhaltselements *tatsächlich wahrgenommen* für dessen Zutreffen diskutieren. Solche Erwägungen sind meistens in den Verlauf der Argumentation der jeweiligen Entscheidungsbegründung eingebettet. Da in dieser typischerweise neben Argumenten für eine Entscheidung auch Argumente dagegen untersucht und schlussendlich angenommen oder auch zurückgewiesen werden, können in argumentativ eingebundenen Präzisierungen Merkmale “von der positiven und negativen Seite her” betrachtet werden: Es werden nicht nur Anwendungsbedingungen, sondern auch Ausschlussbedingungen in Betracht gezogen und angenommen, modifiziert oder gegebenenfalls auch abgelehnt. In (2.1) wird so beispielsweise eine in Erwägung gezogene Anwendungsbedingung zunächst in Satz 2 abgewiesen und dann in Satz 3 in einer abgeschwächten Variante akzeptiert.

Dabei geht es in der Regel nur um die für das jeweilige Argument benötigten Festlegungen. Die entsprechenden Äußerungen enthalten daher generell keine vollständigen Definitionen mit hinreichenden und notwendigen Bedingungen. Aufgrund des dialektisch-deliberativen Charakters der juristischen Argumentation kann sich eine Klassifikation solcher argumentativ eingebundenen

Definitionen jedoch nicht nur auf das Merkmalspaar hinreichend / notwendig stützen. Zusätzlich muss berücksichtigt werden, ob die angegebene Information als *Anwendungs-* oder *Ausschluss*bedingung intendiert ist und schließlich, ob eine Bedingung als gültig *festgestellt* oder *abgelehnt* wird. Ausgehend von den zwei möglichen Grundtypen unvollständiger Definitionen ([+hinreichend, -notwendig] und [-hinreichend, +notwendig]) ergeben sich aufgrund dieser Merkmale acht Konstellationen, die in Tabelle 2.2 schematisch zusammengefasst sind. Sprachlich werden diese entweder durch spezielle Prädikate (wie *gleichgültig ist* in (2.1)) oder durch Kombinationen von Negations- und Fokuspartikeln markiert. In Tabelle 2.2 sind ihnen jeweils typische Formulierungsmuster auf der Grundlage des Verbs *vorliegen* zugeordnet (vgl. auch 2.3.2).

| | affirmativ | Annahmebedingung | hinreichend | notwendig | Realisierungsmuster |
|---|---|---|---|---|---|
| I | + | + | + | - | *A liegt insbesondere dann vor, wenn B* |
| II | + | + | - | + | *A liegt nur dann vor, wenn B* |
| III | + | - | + | - | *A liegt insbesondere dann nicht vor, wenn B* |
| IV | + | - | - | + | *A liegt nur dann nicht vor, wenn B* |
| V | - | + | + | - | *A liegt nicht schon dann vor, wenn B* |
| VI | - | + | - | + | *A liegt nicht nur dann vor, wenn B* |
| VII | - | - | + | - | *A liegt nicht schon dann nicht vor, wenn B* |
| VIII | - | - | - | + | *A liegt nicht nur dann nicht vor, wenn B* |

Tabelle 2.2: Grundkonstellationen argumentativ eingebundener unvollständiger Definitionen

Von den 176 definitorischen Sätzen in dem von untersuchten Korpus ließen sich 144 sinnvoll hinsichtlich ihrer Vollständigkeit und argumentativen Funktion klassifizieren, 66 davon spezifizierten zugleich hinreichende und notwendige Bedingungen, sind also im logischen Sinne vollständig (wie bereits erwähnt handelte es sich dabei sämtlich um Kerndefinitionen). Die Verteilung der verbleibenden 78 Sätze mit unvollständigen Definitionen auf die in Tabelle 2.2 beschriebenen argumentativen Konstellationen ist dem Diagramm in Abb. 2.1 zu entnehmen. Die 32 nicht klassifizierbaren Sätze enthielten zu etwa gleichen Tei-

len Angaben von Synonymen, rein extensionale Definitionen durch Aufzählung von Bezugsgegenständen und nicht bedingungsförmige Erläuterungen zum jeweiligen Definiendum (vgl. 2.2.3).

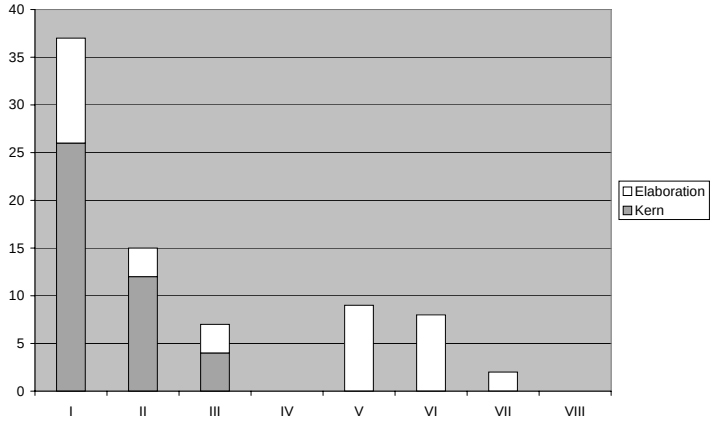



Abbildung 2.1: Verteilung der Definitionen in unserem Korpus auf die argumentativen Typen aus Tabelle 2.2

Kerndefinitionen finden sich in unserem Korpus nur in den affirmativen Konstellationen I, II und III, wobei der Anteil an den Ausschlussbedingungen (III) geringer ist als an den Annahmebedingungen. Die elaborierenden Angaben verteilen sich dagegen etwa zur Hälfte auf die affirmativen und ablehnenden Aussagetypen. Uninstantiiert sind in unserem Korpus lediglich die besonders komplexen Typen IV und VIII. Es ist zu vermuten, dass diese auch generell selten auftreten, da ihnen aufgrund ihrer logischen Struktur für den positiven Anwendungsbereich des jeweiligen Definiendum die geringste Relevanz zukommt.

Die affirmativen Aussagetypen I - IV haben in der Argumentation eine *deduktiv-begründende* Funktion. Als Begriffsbestimmungen sind sie in der Regel ergiebiger als ablehnende Feststellungen. Bei Vorliegen hinreichender Anwendungsbedingungen (Typ I) oder Fehlen notwendiger Ausschlussbedingungen (Typ IV) ist ein Terminus definitiv anwendbar, bei Fehlen notwendiger

Anwendungsbedingungen (II) oder Vorliegen hinreichender Ausschlussbedingungen (III) kann er definitiv abgesprochen werden.

Die argumentative Hauptfunktion der ablehnenden Aussagetypen V-VIII ist dagegen *defensiv*. Sie dienen der Ausräumung tatsächlich (zum Beispiel von einer der Konfliktparteien oder der Vorinstanz) vorgebrachter oder auch vom Gericht vorweggenommener Einwände gegen das verfolgte Argumentationsziel. Im Hinblick auf die Anwendbarkeit eines Begriffs ermöglichen sie (zumindest in Isolation) keine definitive Entscheidung. Trotzdem können sie für die Begriffsverwendung auch jenseits des spezifischen Argumentationszusammenhangs relevant sein, weil die abgewiesenen Anwendungs- oder Ausschlussbedingungen in aller Regel nicht zufällig gewählt sind. Meistens betreffen sie wie in (2.1) naheliegende Gegengründe zu der jeweils getroffenen Entscheidung, so dass damit zu rechnen ist, dass die abgewiesene Bedingung regelmäßig wieder als Einwand gegen die Verwendung des Definiendum vorgebracht werden wird. Wie in unserem Beispiel wird zudem oft im unmittelbaren Kontext (oder auch kontrastierend in derselben Aussage) Information "nachgeliefert", die eine definitive Entscheidung über die Anwendung des Definiendum ermöglicht.

Außerdem weisen nicht alle ablehnenden Aussagen eine Anwendungsbedingung insgesamt zurück. Wie in (2.4) kann auch nur der Status eines Merkmals korrigiert werden (hier: *Hässlichkeit* von notwendig zu hinreichend). Der Aussage kann dann einerseits noch immer direkt eine positive Verwendungsregel entnommen werden (hier: *Hässlichkeit reicht zur Feststellung einer Verunstaltung aus*), andererseits der Hinweis, dass sich eine notwendige Bedingung aus einer Abschwächung des betreffenden Merkmals ergeben könnte.

(2.4) Insoweit setzt die Feststellung einer Verunstaltung kein so krasses geschmackliches Unwerturteil wie das einer das ästhetische Empfinden verletzenden Häßlichkeit voraus.

(Oberverwaltungsgericht Berlin 2. Senat, 31. Juli 1992, AZ 2 B 14.90, juris)

### 2.2.2 Angaben fallbezogener Anwendungsbedingungen

Einen zweiten Haupt-Defintionstypus (dem in erster Linie elaborierende Definitionen zuzuordnen sind) stellen *Angaben fallbezogener Anwendungsbedingungen* dar. Die fallbezogene Ausrichtung solcher elaborierenden Definitionen führt dazu, dass sie nur für einen eingeschränkten Sachbereich Gültigkeit haben. So wird in (2.5), (2.6) und (2.7) eine Anwendungsbedingung (in den Beispielen unterstrichen) angeführt, die die Semantik des Definiendum (fett gedruckt) nicht im allgemeinen, sondern nur für eine bestimmte Klasse von Fällen fest-

legt, zu der jeweils auch der zu entscheidende Sachverhalt gehört. Für die Anwendung des Definiendum in anderen Fällen sind die angegebenen Bedingungen hingegen nicht unbedingt notwendig und hinreichend.

(2.5) Eine nicht unterzeichnete Widerspruchs- oder Klageschrift genügt dann dem Erfordernis der **Schriftform** (vgl. §§ 70 Abs. 1 Satz 1, 81 Abs. 1 Satz 1 VwGO), wenn sich aus ihr allein hinreichend sicher, ohne Rückfrage oder Beweiserhebung ergibt, dass sie von dem Widersprechenden herrührt und mit dessen Willen in den Verkehr gelangt ist.

(Oberverwaltungsgericht für das Land Nordrhein-Westfalen 14. Senat, 25. Juni 2001, AZ 14 A 782/00, juris)

(2.6) Wenn nicht jedes einzelne zu einer Mehrheit von baulichen Anlagen gehörende Bauwerk die Voraussetzungen des Art. 1 Abs. 1 DSchG erfüllt, liegt ein **Ensemble** nur vor, wenn das gegenwärtig vorhandene Orts-, Platz- oder Straßenbild insgesamt erhaltenswürdig ist.

(BayObLG 3. Senat für Bußgeldsachen, 3. März 1993, AZ 3 ObOWi 17/93, juris)

(2.7) Bei einem Einfamilien-Reihenhaus liegt ein **mangelhafter Schallschutz** dann vor, wenn die Haustrennwand einschalig errichtet wurde und der Schalldämmwert von 57 dB nicht erreicht wird.

(OLG Stuttgart 1. Zivilsenat, 22. November 1995, AZ 1 U 199/93, juris)

Die drei betrachteten Äußerungen unterscheiden sich im Hinblick auf die Relation zwischen der angegebenen elaborierenden Bedingung und der vollen Semantik des Definiendum. In den ersten beiden Beispielen, (2.5) und (2.6), wird eine abschwächende bzw. verstärkende Modifikation der allgemeinen Semantik des Definiendum für gewisse Sonderfälle vorgenommen: (2.5) schwächt die allgemein gültige Anforderung für die Verwendung des Terminus *Schriftform* (nämlich *Vorhandensein einer eigenhändigen Unterschrift*) im vorgegebenen Kontext ab, (2.6) formuliert eine (verstärkende) Zusatzanforderung für die Verwendung des Begriffs *Ensemble* in den durch die Vorbedingung charakterisierten Fällen. Im dritten Beispiel (2.7) dagegen stellt die angegebene Bedingung im Vergleich zur allgemeinen Semantik des Definiendum keine Abschwächung oder Verstärkung dar. Sie enthält vielmehr eine Präzisierung: Die Einschränkung auf eine klar abgrenzbare Klasse typischer Anwendungsfälle (nämlich

*Einfamilienhäuser*) erlaubt die Erfassung einer wertenden Begriffskomponente (*mangelhaft*) durch einen quantitativen Grenzwert.

Der Sachbereich, für den fallbezogene Elaborationen gültig sind, bleibt oft implizit und muss dem Kontext entnommen werden. In (2.5), (2.6) und (2.7) wird er jedoch auf unterschiedliche Weise in der Definition selber angegeben. In (2.5) wird das Definiendum *Schriftform* in einen Verwendungskontext gestellt, der dem zu entscheidenden Fall entspricht. In (2.6) und (2.7) wird dagegen der Anwendungsbereich durch eine eigene Angabe bestimmt. In (2.6) geschieht dies durch eine als Konditionalsatz realisierte Vorbedingung für die Anwendung der Definition (in der gesamten Äußerung werden somit zwei Bedingungen – Vorbedingung und Definiens – für die Verwendung des Definiendum aufgestellt). In (2.7) benennt eine *bei*-Präpositionalphrase eine Klasse von Gegenständen, auf die das Definiendum potentiell anwendbar ist und für deren Einordnung das angegebene Definiens gelten soll.

Einschränkungen durch den Verwendungskontext (wie in (2.5), wo die Definition des Begriffs *Erfordernis der Schriftform* kontextuell auf den Fall *nicht unterzeichneter Widerspruchs- und Klageschriften* eingeschränkt wird) finden sich oft (aber nicht ausschließlich) bei Definitionen kontextabhängiger Begriffe, für deren Bestimmung aus semantischen und meist auch syntaktischen Gründen ohnehin ein Bezugsausdruck erforderlich ist. Für die Wahl zwischen den in (2.6) und (2.7) illustrierten Einschränkungstypen scheinen hingegen vor allem stilistische Gesichtspunkte eine Rolle zu spielen.[3]

In welchem Maße fallbezogene Anwendungsbedingungen für die Verwendung des definierten Terminus von genereller Relevanz sind, hängt vor allen Dingen davon ab, ob sie eine Übertragung auf weitere Fälle zulassen oder nicht. Dafür kommt es darauf an, wie eng die mit der jeweiligen Feststellung verknüpfte Einschränkung des Geltungsbereichs gefasst ist. Auch aus der angegebenen Verwendungsbedingung selber kann sich jedoch unter Umständen – wie in (2.8) – eine so enge Einschränkung ergeben, dass mit hoher Wahrscheinlichkeit nur ein seltener Einzelfall erfasst wird.

(2.8) Wird ein öffentlicher Weg unpassierbar, weil Hecken und Gebüsch ihn überwuchern, so liegt eine Gefahr für die öffentliche Sicherheit und Ordnung vor [...]

(OLG Frankfurt, 10. August 1982, AZ 3 Ws (B) 141/82 OWi, juris)

[3] So ist in (2.6) eine detaillierte Beschreibung des für die Anwendbarkeit der Definition vorausgesetzten Szenarios erforderlich. Diese kann in einem Nebensatz durch weniger komplexe Konstruktionen ausgedrückt werden als in einer Präpositionalphrase. In (2.7) steht dagegen mit dem Ausdruck *Einfamilienreihenhaus* ein Domänenterminus zur kompakten Bezeichnung des Anwendungsbereichs zur Verfügung.

### 2.2.3 Kommentierende und erläuternde Zusatzinformation

Ein dritter Haupttypus (wiederum vor allem elaborierender) definitorischer Aussagen umfasst verschiedene Arten von kommentierenden und erläuternden Angaben, deren Gemeinsamkeit darin besteht, dass die in ihnen enthaltene Information zwar selbst nicht direkt zur Begriffsbestimmung beiträgt (also insbesondere nicht direkt argumentativ oder deduktiv verwertbar ist), jedoch für die Interpretation und Anwendung von Angaben der bisher besprochenen Arten relevant ist.

Eine häufig auftretende Art der Beziehung zwischen definitorischer Information im engeren Sinne und solchen Zusatzangaben illustriert die Textpassage in (2.9). Die zitierten Sätze schließen sich im Quelltext an die elaborierende Definition in (2.3) an. Sie beziehen sich also auf den Begriff *Verunstaltung*.

(2.9) (1) Maßgebend ist dabei das Empfinden des gebildeten Durchschnittsbetrachters, d.h. eines für ästhetische Eindrücke offenen, jedoch nicht besonders empfindsamen und geschulten Betrachters. [...]

(2) Ob eine Werbeanlage eine solche Wirkung hervorruft, ist unter Berücksichtigung der gesamten Umstände des Einzelfalls zu beurteilen, wobei auch die Funktion des jeweils betroffenen Baugebiets zu berücksichtigen ist (vgl. VGH Baden-Württemberg, Urteil vom 07.08.1986 - 8 S 994/86 - mit weiteren Nachweisen).

(Verwaltungsgerichtshof Baden-Württemberg 5. Senat, 12. August 1993, AZ 5 S 1018/92, juris)

Die Definition (2.9) wird hier um weitere Information zum selben Begriff ergänzt, die aber nicht die Semantik des Definiendum bestimmt, sondern auf die korrekte Vorgehensweise bei den für die Anwendung notwendigen Erwägungen hinweist.

Neben solchen Angaben zu Abwägungen, die an den Begriffsanwender gerichtet sind, finden sich als erläuternde Zusatzinformation beispielsweise:

- illustrierende Beispiele
- Maximen zum Umgang mit Interessenkonflikten
- reflektierende Überlegungen zum Spielraum und der Gewichtung der Methoden bei der Auslegung des Definiendum (z.B. *ist weit auszulegen* oder *wird vom Gesetzgeber nicht bestimmt*)

- zusammenfassende allgemeine Schlussfolgerungen aus Definitionen (insbesondere Hinweise zu den rechtssystematischen Konsequenzen einer bestimmten Begriffsverwendung)

Für die hier genannten Informationstypen ist kaum allgemein zu sagen, bis zu welchem Punkt eine Aussage relevant für die semantische Bestimmung eines Terminus ist. Auch hier spielt eine wichtige Rolle, in wie weit die jeweilige Aussage eine Übertragung auf andere Fälle zulässt. Allerdings zeigt zum Beispiel (2.9), dass als generalisierende Aussagen unter Umständen relativ inhaltsarme "eingeschliffene" Ausdrücke Verwendung finden, die in der juristischen Literatur mitunter als *Leerformeln* bezeichnet werden. Diese enthalten – wie *gebildeter Durchschnittsbetrachter* in Satz 1 des Beispiels – deutlich weniger verwertbare Information als stärker auf den Einzelfall ausgerichtete Feststellungen (vgl. Satz 2). Die besonders ausgeprägte Problematik der Abgrenzung zu Nicht-Definitionen schlägt sich auch darin nieder, dass von den 32 in unserer Annotationsstudie als *erläuternd bzw. kommentierend* zu klassifizierenden Sätzen ein besonders hoher Anteil (6 Sätze, also etwa 19%) nur den Konfidenzwert *zweifelhaft* erhielten.

### 2.2.4 Fazit

Der Unterschied zwischen Definitionen und anderen Aussagen kann in Urteilsbegründungen also nicht strikt an den semantischen Kriterien der Äquivalenz und Essentialität festgemacht werden.

Das Essentialitätskriterium bleibt zwar auch für Definitionen in dem hier diskutierten weiten Sinne insofern bestehen, als rein kontingente, z.B. vollständig auf empirische Beobachtungen gestützte Aussagen generell nicht als Definitionen in Frage kommen, selbst wenn sie wie (2.10) klar einem der definionstypischen Formulierungsschemata entsprechen (im Beispiel dem aristotelischen Schema *Ein A ist ein B, das C*).

(2.10) Typische Beispiele sind Lagerhäuser und -plätze, die als selbständige Gewerbebetriebe mit dem alleinigen oder zumindest überwiegenden Nutzungszweck der Lagerung geführt werden.

(BVerwG 4. Senat, 8. November 2001, AZ 4 C 18/00, juris)

Das Äquivalenzkriterium kann hingegen zumindest für elaborierende Aussagen nicht aufrecht erhalten werden. Wie die oben betrachteten Beispiele gezeigt haben, ist der Übergang zwischen definitorischen und nicht-definitorischen Aussagen hier fließend und basiert auf einem Bündel graduierbarer, teils der semantischen, teils eher der pragmatischen Ebene zuzuordnender Merkmale.

Zusammenfassend lassen sich aus der bisherigen Diskussion zwei Leitfragen ableiten, die als Grundlage einer operationalisierten Unterscheidung zwischen Definitionen und Nicht-Definitionen in Entscheidungsgründen dienen können:

1. Befasst sich eine Aussage mit einem bestimmten Sprachgebrauch bzw. stützt sie sich für ihre Geltung auf eine Norm darüber? Diese Fragestellung dient zur Ausgrenzung rein kontingenter Aussagen und setzt somit das Essentialitätskriterium in der eben erwähnten Form um.

2. Erlaubt der Verallgemeinerungsgrad der Aussage die Anwendung auf eine größere Zahl verschiedener Fälle? Insbesondere sind Aussagen, in denen auf den zu entscheidenden Einzelfall Bezug genommen wird (Subsumtion im engeren Sinne) normalerweise nicht als Definitionen anzusehen. Dasselbe gilt für Aussagen, die zwar der Formulierung nach verallgemeinern, aber ein so spezielles Definiens enthalten, dass sie im Grunde einen Einzelfall charakterisieren (wie (2.8)).

Bei der Beantwortung der ersten Frage bieten in der Regel sprachliche Merkmale wichtige Anhaltspunkte. Ein solches Merkmal ist z.B. das Vorhandensein eines Definitionsprädikats (vgl. 2.3.2). Auch für die zweite Frage sind sprachliche Indikatoren relevant, z.B. die verwendeten Determinatoren und der Abstraktionsgrad der im Definiens gebrauchten Begriffe. Eindeutig zu beantworten ist sie jedoch oft nur auf der Grundlage von speziellem Fach- und Domänenwissen, da typischerweise entschieden werden muss, ob das Definiendum Tatbestandsmerkmal einer Norm ist oder zumindest bei der Auslegung eines solchen Tatbestandsmerkmals eine Rolle spielen kann, und ob das Definiens Fälle betrifft, die im Regelungsbereich der betreffenden Norm mit einer gewissen Plausibilität wiederholt auftreten können.

Im Rest diese Kapitels befassen wir uns zunächst mit sprachlichen Merkmalen richterlicher Definitionen. Wir untersuchen, mit welchen Ausdrucksmitteln die in unserem Korpus vorgefundenen Definitionen markiert sind und wie sich diese Ausdrucksmittel nach semantisch-funktionalen Gesichtspunkten einteilen lassen (2.3).

Im Anschluss diskutieren wir die Ergebnisse einer zweiten Korpusstudie, in der wir mittels einer doppelten Annotation ermittelt haben, welches Maß an intersubjektiver Übereinstimmung bei der Identifikation von Definitionen auf der Basis der oben angeführten Leitfragen erzielbar ist (2.4).

Unsere Studie fokussiert damit insgesamt im wesentlichen die erste der beiden oben genannten Leitfragen. Soweit wir im weiteren Verlauf unserer Arbeit Aussagen zur zweiten Fragestellung machen, stützen wir uns auf Einzelbeobachtungen aus unseren Korpora. Die Ergebnisse bleiben also zu ergänzen um

eine Systematisierung der sprachlichen Mittel zur Regulierung der Allgemeingültigkeitsanspruchs richterlicher Definitionen. Für diese dürfte allerdings ein erheblich höheres Maß an juristischer Fachkompetenz nötig sein als für die hier diskutierten Untersuchungen.

## 2.3 Sprachliche Form

In diesem Abschnitt analysieren wir die sprachlichen Mittel, die zum Ausdruck von Definitionen in Urteilsbegründungen verwendet werden. Damit legen wir zugleich die Grundlage für den in den folgenden Kapiteln dieser Arbeit beschriebenen Ansatz zur automatischen Extraktion solcher Definitionen aus Urteilstexten anhand sprachlicher Merkmale.

### 2.3.1 Bisherige Untersuchungen zur Form von Definitionen

Wie in 1.2.1 erwähnt wurden Definitionen lange Zeit vor allem in der philosophischen Literatur diskutiert, wobei vor allem ontologische und erkenntnistheoretische Fragestellungen im Vordergrund standen. Die sprachliche Form von Definitionen thematisieren erst in neuerer Zeit Studien in der angewandten Sprachwissenschaft, vor allem in der Terminologiewissenschaft und Lexikographie.

Als typische Ausdrucksform für Definitionen wird in der terminologiewissenschaftlichen und lexikographischen Literatur meistens weiterhin das auf Aristoteles zurückgehende Schema der Definition durch Angabe des *genus proximum* und der *differentia specifica* des zu definierenden Begriffs angesehen, also seines nächstgelegenen Oberbegriffs und des Merkmals, durch das sich der Begriff von den anderen Unterbegriffen desselben Oberbegriffs unterscheidet (vgl. für die Lexikographie z.B. (Landau, 2001, 138ff.) und (Ilson, 1986, 218ff.) und für die Terminologiewissenschaft Arntz u. a. (2004), (Strehlow, 1983, 20) und die bereits in 1.2.1 erwähnten terminologischen Grundsatznormen). Als sprachliches Grundmuster für solche Definitionen wird häufig die folgende Formel angeführt:

> Ein *A* ist ein *B*, das *C*

Terminologiewissenschaftliche und lexikographische Untersuchungen zu Definitionen verfolgen meist eine präskriptive oder didaktische Zielsetzung. Es wird dargestellt, mit welchen Mitteln in einem bestimmten Bereich Definitionen getroffen werden *sollen*, um ihren Zweck möglichst gut zu erfüllen (z.B. die Systematisierung von Fachsprachen, die Unterstützung des Sprachlernenden bei Produktion und / oder Rezeption oder die Nutzung als Referenz für den

Muttersprachler in Zweifelsfällen). In diesem Zusammenhang ist der Aufbau von Definitionen nach der dargestellten Formel sinnvoll, weil die Grundoperationen Äquivalenzbildung, Klassifikation und Einschränkung besonders transparent zum Ausdruck kommen: Die Kopulakonstruktion stellt die Äquivalenz zwischen dem Definiendum *A* und dem klassifikatorischen Genusbegriff *B*, eingeschränkt durch die als Relativsatz ausgedrückte differentia specifica *C*, her.

Die natürliche Sprache lässt jedoch für denselben Inhalt fast immer unterschiedliche, semantisch weitgehend äquivalente Realisierungen zu. Zwischen diesen trifft der Sprecher eine Auswahl. Diese mag zwar teils auf Transparenzgesichtspunkten beruhen. Sie wird aber auch durch ein Reihe weiterer Faktoren, etwa den Kontext, Anforderungen des Genres und nicht zuletzt persönliche stilistische Vorlieben beeinflusst. Die Frage, mit welchen unterschiedlichen Formulierungen Definitionen als Resultat all dieser Einflüsse in verschiedenen Textsorten tatsächlich getroffen werden, ist in der Terminologiewissenschaft bisher nur vereinzelt ausführlicher untersucht worden. Erwähnenswert sind vor allem die Untersuchungen von Trimble (1985), Flowerdew (1992), Pearson (1998), Meyer (2001) und Barnbrook (2002).

Barnbrook untersucht die Struktur der Definitionen in *Collins Cobuild Student's Dictionary*, befasst sich also mit einer Textsorte, die nur bedingt als frei formuliert betrachtet werden kann[4] und außerdem weitgehend dekontextualisierte Definitionen enthält. Auf seine Studie werden wir aus diesen Gründen hier nicht näher eingehen. Die in Trimble (1985) dargestellten Ergebnisse wurden nicht systematisch anhand eines Korpus gewonnen sondern beruhen auf Einzelbeobachtungen in verschiedenen wissenschaftlich-technischen Fachtexten. Meyer legt für ihre Untersuchung ein Korpus mit technischem Text zu Grunde, das sie aber nicht genauer beschreibt. Flowerdew verwendet ein Korpus von insgesamt 329 Definitionen aus transkribierten Biologie- und Chemievorlesungen. Pearson betrachtet Definitionen in Texten aus drei verschiedenen, deutlich größeren Korpora, die sie unterschiedlichen Kommunikationsszenarien zuordnet: Didaktisch angelegte Texte aus einem Lehrbuchkorpus (*expert-to-uninitiated*-Kommunikation, 1 Million Token), technischer Text aus dem *International Telecommunications Union Handbook* (*expert-to-initiate*-Kommunikation, 4,7 Millionen Token) und wissenschaftliche Artikel aus einem Jahrgang der Zeitschrift *Nature* (*expert-to-expert*-Kommunikation, 230 000 Token).

[4] Für die Definitionen in den *Cobuild*-Wörterbüchern wurde ein eigener, an alltagssprachlichen Formulierungsmustern orientierter Stil als Leitlinie entwickelt (vgl. Hanks (1987)). Indes bleiben die generellen Rahmenbedingungen (v.a. Platzknappheit) und besonderen Zielsetzungen (z.B. Vermittlung grammatikalischer Information, Einheitlichkeit und strukturierter Aufbau der einzelnen Artikel) von Wörterbüchern ein bestimmender Einflussfaktor.

In allen vier Untersuchungen dient das Schema der Begriffsbestimmung *per genus et differentiam* als Bezugspunkt für die Klassifikation von Definitionen. Trimble unterscheidet nach ihrem Verhältnis zu diesem Schema *formale*, *semi-formale* und *nicht-formale* Definitionen. In formalen Definitionen sind sowohl Genusterm als auch differentia realisiert, in semi-formalen Definitionen fehlt der Genusterm. *Nicht-formale* Definitionen bestimmt Trimble – ohne genauer auf ein Abgrenzungskriterium zu den beiden anderen Klassen einzugehen – anhand ihrer Aufgabe: *to define in a general sense so that a reader can see the familiar element in whatever the new term may be* (Trimble, 1985, 78). Dafür werde in vielen Fällen ein Synonym (nach Trimble die häufigste Art nicht-formaler Definitionen) oder zumindest ein mit dem Definiendum annähernd gleichbedeutender Ausdruck angegeben.

Alle erwähnten Studien unterscheiden zudem in Definitionen drei sprachliche Bestandteile:

1. Das Definiendum,
2. das Definiens und
3. Material, das diese beiden Bestandteile verknüpft.

Sie betrachten verschiedene Möglichkeiten für die Realisierung des dritten Bestandteils, also der Verknüpfung zwischen Definiens und Definiendum. Nur Pearson benennt allerdings tatsächlich eine größere Anzahl von Ausdrücken, die in dem von ihr untersuchten Korpus in dieser Funktion auftreten.

Pearson übernimmt die von Trimble vorgeschlagene Unterscheidung zwischen formalen, semi-formalen und nicht-formalen Definitionen, trennt jedoch zugleich zwischen *Verb-vermittelten* und *nicht Verb-vermittelten* Definitionen. Sie ordnet dabei den ersten beiden von Trimble unterschiedenen Definitionstypen (*formal* und *semi-formal*) bestimmte definitionsvermittelnde Verben zu, während sie den dritten (*nicht-formal*) mit den durch andere Ausdrücke vermittelten Definitionen identifiziert. Die einzigen *connnective verbs* in den formalen Definitionen in ihren Korpora sind nach Pearsons Angaben die Verben *be*, *consist*, *define*, *denote*, *designate*, *comprise*, *be called*, *be defined (as)* und *be known (as)* (Pearson, 1998, 140). Für die verknüpfenden Ausdrücke in Definitionen der anderen beiden Typen gibt sie keine vollständigen Listen, sondern nur Beispiele an. Als *connective verbs* in semi-formalen Definitionen nennt sie *contain*, *have*, *be used (for)*, *include*, *involve*, *be characterized (by)*, *be described (as)*, *produce* und *provide* (Pearson, 1998, 158), als typische *connective phrases* in nicht-formalen Definitionen parenthetische Klammern und andere Interpunktionszeichen sowie die Ausdrücke *i.e.*, *e.g.*, *called* und *known as* (Pearson, 1998, 169ff.).

Während Pearson wie Trimble annimmt, dass der Grad der Übereinstimmung mit dem aristotelischen Definitionsschema auch die inhaltliche Qualität einer Definition bestimmt, gehen Flowerdew und Meyer nicht davon aus, dass formale Definitionen zwangsläufig informativer als nicht-formale Definitionen sind. Flowerdew stellt fest, dass es sich bei Genusbegriffen in formalen Definitionen oft um Wiederholungen von Bestandteilen eines komplexen Definiendum oder um Ausdrücke mit rein referentieller Funktion handelt (gemeint sind damit zu Relativpronomen korrelative Demonstrativa, im Deutschen also z.B. *dasjenige*, und Relativpronomina zur Einleitung freier Relativsätze, im Deutschen also *was* und *wer*). Meyer weist darauf hin, dass sowohl die Wahl des Genusbegriffs als auch die Wahl der *differentiae* aufgrund der Perspektive des Autors variieren kann, und dass häufig zu weite generische Oberbegriffe anstatt taxonomisch sinnvollerer spezieller Termini als Genusbegriffe angegeben werden. Sie geht außerdem auf Gründe ein, aus denen vollwertige Definitionen ohne Angabe eines Genusbegriffs getroffen werden können. Zum einen vermutet sie, dass die Hyperonymierelation nicht in allen Fachgebieten gleich wichtig zur Begriffsbestimmung ist (als Beispiel nennt sie medizinische Texte, in denen es eher wichtig sei, Krankheiten durch Symptome und Behandlungsmethoden zu charakterisieren, als sie als *Krankheit* einzuordnen). Der Forschung zu Wortnetzen entnimmt sie zum anderen dass Ergebnis, dass die Hyperonomierelation generell nicht für alle Klassen von Substantiven und vor allem nicht für alle Wortarten gleichermaßen zur Bedeutungsbestimmung geeignet sei.

### 2.3.2 Definitorische Formulierungen in Urteilsbegründungen

Auch in Urteilsbegründungen wird – die bisher betrachteten Beispiele lassen dies bereits deutlich erkennen – eine große Zahl verschiedener Formulierungen für Definitionen verwendet. Wie wohl überhaupt beim Verfassen juristischer Texte ist das Streben nach einfacher Verständlichkeit und semantischer Transparenz in Urteilstexten allenfalls einer von vielen Faktoren, die die Formulierungswahl beeinflussen. Bei der Formulierung richterlicher Definitionen treten neben rhetorischen und stilistischen Gesichtspunkten die Orientierung an gängigen Textbausteinen, die Einpassung in den jeweiligen Kontext sowie die möglichst “wasserdichte” Absicherung gegen unbeabsichtigte Schlussfolgerungen aus der getroffenen Festlegung als wichtige Einflüsse hinzu.

#### (a) Übertragbarkeit der Ergebnisse auf rechtssprachliche Definitionen

Auch in Urteilstexten finden sich Definitionen, die einen Begriff “mustergültig” taxonomisch einordnen und abgrenzen (so zum Beispiel die auch in Urteilsbegründungen im gleichen Wortlaut auftretende Definition des Verbraucherbe-

griffs aus §13 BGB, vgl. (1.2.2)). Das aristotelische Definitionsschema spielt aber in diesem Genre nicht die prominente Rolle, die ihm in der Terminologiewissenschaft generell zugewiesen wird. Ein Grund hierfür liegt darin, dass im Rahmen der in den vorigen Abschnitten ausführlicher besprochenen diskursiven Ausdifferenzierung von Begrifflichkeiten meistens nicht die taxonomische Einordnung eines Begriffs zweifelhaft ist, sondern seine genaue Abgrenzung zu benachbarten Konzepten zur Diskussion steht. Aber auch für umfassende Kerndefinitionen werden aus verschiedenen Gründen oft andere Methoden gewählt als die Definition *per genus et differentiam*. Beispielsweise werden (wie schon in 2.2.2 angesprochen) viele kontextabhängige Termini aus grammatikalischen Gründen kontextuell eingebettet definiert, und Definitionen von Handlungsbegriffen beruhen oft auf einem komplexen Geflecht von (unter Umständen zusätzlich mit impliziten Voraussetzungen verknüpften) Beschränkungen über mögliche Handlungsbeteiligte. So gibt die Definition in (2.11) zur Bestimmung des verbalen Terminus *entledigen* nicht einfach Merkmale des übergeordneten Handlungsbegriffs *Gewahrsamsaufgabe* an, sondern eine komplexe Beschränkung, die dessen Agens- und Patiensrolle zueinander in Relation setzt (Intentionen des Besitzers hinsichtlich der weiteren Verwendung der besessenen Sache) und aus der wiederum Information über die vorausgesetzte Situation (*Jemand besitzt eine bewegliche Sache*) entnommen werden kann.

(2.11) Entledigen bedeutet, daß der Besitzer den Gewahrsam an der beweglichen Sache aufgibt, ohne damit zugleich einen anderen Zweck im Sinne einer irgendwie gearteten weiteren Verwendung der Sache (wirtschaftliche Verwertung, Verschenken o.ä.) zu verfolgen (. . . ).

(BVerwG 7. Senat, 19. Dezember 1989, AZ 7 B 157/89, juris)

Auch Pearsons Unterscheidung zwischen durch *connective verbs* und *connective phrases* vermittelten Definitionen ist zwar grundsätzlich auf Definitionen in Urteilsbegründungen übertragbar, läßt aber keine direkten Rückschlüsse auf deren inhaltliche Qualität zu. So basiert in den folgenden Beispielen die Definition des Terminus *Nachfluchtgründe* in (2.12) auf dem "Definitionsverb" *vorliegen*, in (2.13) ist sie durch parenthetische Klammern und den Signalausdruck *sog.* markiert.

(2.12) Ein Nachfluchtgrund liegt vor, wenn dem Asylbewerber aufgrund von Umständen, die nach seiner Ausreise aus seinem Heimatland eingetreten sind, bei einer Rückkehr dorthin jetzt eine politische Verfolgung mit beachtlicher Wahrscheinlichkeit droht.

(Hessischer Verwaltungsgerichtshof 10. Senat, 30. Januar 1995, AZ 10 UE 2626/92, juris)

(2.13) Fluchtgründe, die erst mit oder nach Verlassen des Herkunftslandes begründet werden (sog. Nachfluchtgründe), vermögen daher regelmäßig nur Abschiebungsschutz über § 51 Abs. 1 AuslG und nur in besonderen Fällen das Asylgrundrecht des Art. 16a Abs. 1 GG zu vermitteln, wenn sog. beachtliche Nachfluchtgründe vorliegen.

(Sächsisches Oberverwaltungsgericht 4. Senat, 28. August 2001, AZ A 4 B 4388/99, juris)

Trotz des Unterschieds in der Formulierung enthalten beide Sätze jedoch inhaltlich weitgehend gleichwertige Festlegungen.[5]

Im Folgenden gehen wir zunächst auf die in unserer Korpusstudie identifizierten Varianten *connective phrase*-vermittelter Definitionen (wir verwenden die Bezeichnung *parenthetische und appositive Definitionen*) ein, dann auf die verschiedenen Formulierungsmuster *connective verb*-vermittelter (oder allgemeiner gesprochen *prädikatbasierter*) Definitionen. Für diese Kategorie schlagen wir eine genauere Untergliederung vor. Im Anschluss untersuchen wir die Realisierungsmöglichkeiten von Definiens und Definiendum sowie weiteren Informationseinheiten in prädikatbasierten Definitionen.

#### (b) Appositive und parenthetische Definitionen

Die von Pearson als Beispiele für definitionsvermittelnde *connective phrases* angeführten Ausdrücke dienen ihrer linguistischen Funktion nach zur Einleitung oder zusätzlichen Markierung parenthetischer und appositiver Elemente, während die von ihr genannten *connective verbs* in Definitionen als Prädikat fungieren. Die Einteilung in *connective verbs* und *connective phrases* reflektiert somit keine Differenz im Informations*gehalt* von Definitionen, sondern einen Unterschied in der Informations*struktur* des jeweiligen Kontexts: Unter Verwendung definitionsvermittelnder Prädikate (entscheidend ist auch hier nicht die Wortart, sondern die grammatikalische Funktion des verknüpfenden Ausdrucks in der jeweiligen Äußerung) wird eine Definition als Hauptaussage eines Satzes vorgebracht. In appositiven und parenthetischen Definitionen wird hingegen das Definiendum oder Definiens als Zusatzinformation in einen Satz eingeschoben, in dem der jeweils andere Definitionsbestandteil bereits eine von

[5] Zudem spezifiziert die *connective phrase*-basierte Definition in (2.13) einen Oberbegriff (*Fluchtgründe*), der durch einen Relativsatz modifiziert wird. Sie enthält somit bis auf die Kopula alle Elemente einer aristotelischen Definition. Daher wäre sie nach der Einteilung Trimbles sinngemäß wohl eher als *formal* denn als *nicht-formal* einzuordnen. Auch in der (nach Trimble als *semi-formal* zu klassifizierenden) Definition in (2.12) sind alle Sinnelemente des Genusbegriffs realisiert, allerdings in ausformulierter und somit weniger kompakter Form, nämlich durch die Anforderung, dass *dem Asylbewerber eine politische Verfolgung mit beachtlicher Wahrscheinlichkeit droht*.

der Definition unabhängige Funktion erfüllt. Dadurch tritt auch die Definition insgesamt in den Hintergrund gegenüber der Hauptfunktion des Satzes. So erhält die Definition in (2.12) durch die prädikatbasierte Realisierung einen hervorgehobenen Status, während in (2.13) das Definiens durch appositive Realisierung gegenüber den anderen Inhalten des Satzes in den Hintergrund gerückt wird. Dadurch kommt auch der Definitionsfunktion gegenüber der Funktion des Satzes als Feststellung (zu den Voraussetzungen für das Zustandekommen eines Abschiebungsschutzes) geringere Prominenz zu.

Außer durch die deutschen Entsprechungen der von Pearson genannten *connective phrases* wird ein solcher Effekt in den von uns untersuchten Texten häufig dadurch erzielt, dass – wie in (2.14) für die Begriffe *objektiver* bzw. *subjektiver Nachfluchtgrund* – das Definiens an das Definiendum als nichtrestriktiver Relativsatz angeschlossen wird. Ebenso kann es erläuternd in einem eigenen Hauptsatz nach der nicht-definitorischen Hauptaussage realisiert werden, wie die partielle Umschreibung des Terminus *fehlendes Rechtsschutzbedürfnis* in (2.15).

(2.14) Dabei ist zu unterscheiden zwischen objektiven Nachfluchtgründen, die durch Vorgänge im Heimatland des Asylbewerbers unabhängig von seiner Person ausgelöst wurden, und subjektiven Nachfluchtgründen, die der Asylbewerber nach Verlassen des Heimatstaates aus eigenem Entschluss geschaffen hat [...].

(Hessischer Verwaltungsgerichtshof 9. Senat, 28. Februar 2002, AZ 9 UE 1653/98.A, juris)

(2.15) ... fehlt der Beschwerde der Antragsgegnerin das Rechtsschutzbedürfnis. Sie kann ihre Rechtsstellung im Beschwerdeverfahren nicht mehr verbessern.

(Bayerischer Verwaltungsgerichtshof München 22. Senat, 18. Juni 2002, AZ 22 CE 02.815, juris)

Auch in diesen Konstellationen erhält das Definiens und mithin auch die gesamte Definition einen kontextuell zurückgesetzten, kommentarartigen Status.

Tabelle 2.3 gibt einen Überblick über die Formulierungsmuster der parenthetischen und appositiven Definitionen in unserem Korpus. Als *zweifelhaft* annotierte Definitionen sind nicht berücksichtigt.

Aufgrund der in Relation zum Kontext geringeren Prominenz, die einer Definition bei appositiver bzw. parenthetischer Realisierung zukommt, wird dieser Formulierungstyp vor allem verwendet, um als bekannt vorausgesetzte Begriffsbestimmungen aufzugreifen. Appositive und parenthetische Definitionen sind daher als Quelle für neues terminologisches Wissen weniger relevant als

| *Typ* | *Anzahl* | *Zusätzliche Signale* |
|---|---|---|
| *Definiendum in Parenthese* | | |
| in Klammern | 10 | *sogenannt* |
| durch andere Interpunktionszeichen abgetrennt | 2 | *und damit, also, sogenannt* |
| *Definiens in Parenthese* | | |
| als Einschub im selben Satz (in Klammern, Kommata oder Gedankenstrichen) | 6 | *das heißt, also, als* |
| als Relativsatz | 8 | |
| als eigener Satz | 6 | *danach, dann* |

Tabelle 2.3: Parenthetische Definitionen

prädikatbasierte Definitionen. Zudem sind sie (wie aus Tabelle 2.3 zu entnehmen ist) sprachlich nur sehr unspezifisch markiert und bieten daher nur wenige Anhaltspunkte für eine trennscharfe automatische Erkennung und Verarbeitung.[6] Wir werden uns aus diesen Gründen im Rahmen dieser Arbeit nicht mehr näher mit appositiven und parenthetischen Definitionen befassen.

#### (c) Definitionsvermittelnde Prädikate

Die in unserem Korpus in prädikatbasierten Definitionen verwendeten Prädikate stimmen nur zu einem relativ geringen Teil mit den von Pearson genannten *connective verbs* überein. Insgesamt ist ihre Zahl größer und es finden sich semantisch sehr unterschiedliche Ausdrücke. Diese lassen sich jedoch relativ eindeutig danach einteilen, ob die mit ihnen aufgestellten Definitionen der metasprachlichen oder der objektsprachlichen Ebene zuzurechnen sind. Innerhalb dieser Klassen ist dann eine weitere (zum Teil allerdings weniger eindeutige)

[6] Beispielsweise wäre es allein für die Identifikation von Definitionen der zwei nach Tabelle 2.3 häufigsten Typen nötig, definitorische Klammerverwendungen von anderen Fällen zu unterscheiden sowie nicht-restriktive Relativsätze zu erkennen und auch hier zwischen definitorischen und nicht-definitorischen Fällen zu trennen. Diese Aufgaben sind wohl nur auf der Grundlage eines weitgehenden Textverstehens zu lösen. Sie dürften daher bis auf weiteres außerhalb der Reichweite praktisch einsetzbarer computerlinguistischer Verfahren liegen.

Einteilung aufgrund semantisch-funktionaler Gesichtspunkte möglich. Es ergeben sich insgesamt fünf verschiedenen Klassen von Definitionsprädikaten.

Metasprachliche Definitionen werden mit Prädikaten vorgenommen, die:

1. explizit die Bedeutungsrelation thematisieren (wie das Verb bedeuten in (2.11)).

2. auf den juristischen Prüfungsvorgang im Zusammenhang mit der Verwendung des jeweiligen Definiendum Bezug nehmen (wie die Prädikate *gleichgültig ist* und *genügt* in (2.1)).

In objektsprachlichen Definitionen werden verwendet:

3. die Kopula *sein* und andere Verben, mit denen eine Klassifikation ausgedrückt wird (z.B. *handeln um*).

4. semantisch leichte Verben (z.B. *vorliegen* oder *gegeben sein*) und mit dem Definiendum assoziierte Funktionsverben (z.B. *genießen* für Rechte), also Verben, durch die ein neutraler Kontext für die prädikative Verwendung des Definiendum geschaffen wird.

5. Verben, die ein Merkmal bezeichnen, durch das das Definiendum in der jeweiligen Definition bestimmt wird (z.B. *Bestandteil* bei dem Prädikat *gehören (zu)*).

Die von Pearson aufgezählten Verben aus formalen Definitionen sind nach dieser Einteilung hauptsächlich der Klasse (1) zuzuordnen, diejenigen aus semi-formalen Definitionen der Klasse (5). In beiden Fällen gibt es jedoch Ausnahmen, z.B. die Verben *comprise* und *consist* in formalen bzw. *be described as* und *be characterized as* in semi-formalen Definitionen.

Die in unserer Pilotstudie identifizierten Definitionsprädikate und ihre Zuordnung zu den eben erläuterten Klassen sind Tabelle 2.4 zu entnehmen (es sind wiederum keine als *zweifelhaft* annotierten Definitionen berücksichtigt).

In Definitionen treten alle der genannten Prädikate im Indikativ Präsens Aktiv bzw., soweit sie im Aktiv ein agentives Subjekt erfordern, passiviert auf. Abweichungen in Tempus und Modus des Definitionsprädikats deuten in den meisten Fällen darauf hin, dass die betreffende Aussage entweder gar keine oder keine verwertbare Definition darstellt. Steht das Prädikat in einem Vergangenheitstempus, handelt es sich in der Regel um eine partikulare Feststellung zum jeweiligen Sachverhalt. Verwendung des Konjunktiv I oder II zeigt an, dass eine Definition entweder angeführt oder theoretisch in Erwägung gezogen, jedenfalls später nicht akzeptiert wird. Dagegen bleibt bei einer Modalisierung durch *können* der definitorische Charakter einer Aussage bestehen, allerdings

| | |
|---|---|
| **(1) Bedeutungsbezogen** | |
| *sprechen (von)* | 2 |
| *bedeuten* | 1 |
| *besagen* | 1 |
| *gekennzeichnet sein* | 1 |
| *umschreiben* | 1 |
| **(2) Prüfungsbezogen** | |
| (a) Vermutungen und Fiktionen | |
| *gelten, ansehen, betrachten, anerkennen (als), annehmen, ausgehen (von)* | 15 |
| (b) Logischer Status | |
| *müssen, erfordern, erforderlich sein, fordern, verlangen* | 14 |
| *voraussetzen, Voraussetzung sein* | 6 |
| *ausreichend sein, genügen* | 3 |
| *entfallen, Ausnahmen geben (von)* | 3 |
| *gleichgültig sein* | 1 |
| (c) Prüfungsmaßstäbe | |
| *entscheidend, maßgebend sein* | 2 |
| *beurteilt, ermittelt werden, zu entnehmen sein* | 3 |
| (d) Extensionale Angaben | |
| *umfassen, einbeziehen (in)* | 2 |
| (e) Beziehungen zwischen institutionellen Begriffen | |
| *folgen (aus), ausschließen* | 3 |
| **(3) Kopula / Klassifikation** | |
| *sein* | 40 |
| *verstehen (unter / als)* | 6 |
| *werden* | 2 |
| *darstellen* | 1 |
| *sich handeln (um)* | 1 |
| *zählen (zu)* | 1 |
| **(4) Semantisch leicht** | |
| *vorliegen* | 8 |
| *gegeben sein* | 4 |
| *der Fall sein* | 1 |
| *genügen (+ Dativ)* | 1 |
| *zukommen* | 1 |
| *haben* | 1 |
| *genießen* | 1 |
| **(5) Merkmal-spezifisch** | |
| *schützen* | 2 |
| *gehören (zu)* | 2 |
| *sich richten (gegen)* | 1 |
| *Gegenstand sein* | 1 |
| *sicherstellen (dass)* | 1 |
| *sich begnügen müssen* | 1 |
| *sich beziehen (auf)* | 1 |

Tabelle 2.4: Klassifikation der Definitionsprädikate im untersuchten Korpus

wird unter Umständen die Kraft oder Reichweite der Definition modifiziert (siehe hierzu 2.3.2).

Im Gegensatz zu Pearsons Beobachtungen finden sich in den von uns untersuchten Dokumenten in Definitionen deutlich weniger Verben, die den Definitionsvorgang selbst thematisieren. Dagegen enthalten sowohl die Klasse der *Merkmal-spezifischen* Prädikate als auch die (bei Pearson nicht vorkommenden) Klassen *semantisch leichte* und *prüfungsbezogene* Prädikate jeweils eine größere Anzahl verschiedener Ausdrücke.

Die (nach Auftreten) größte einzelne Gruppe von Definitionsprädikaten stellen in unserem Korpus *Kopula und Klassifikationsverben* dar. Mit diesen Verben ausgedrückte Definitionen entsprechen in der Regel formal dem *genus et differentia*-Schema, allerdings bestätigen unsere Beobachtungen die Anmerkungen von Flowerdew und Meyer zu den Problemen einer solchen Analyse: In vielen der Definitionen in unserem Korpus ist das angegebene Prädikativum kein taxonomisch sinnvoller Oberbegriff des Definiendum, sondern erfüllt eine rein referentielle Funktion oder ist ad hoc gewählt. Etwa die Hälfte der mit der Kopula *sein* formulierten Definienda bestimmen außerdem ein adjektivisches Definiendum und der "Genusbegriff" ist als Bezugswort aus syntaktischen Gründen notwendig.

Die zweitgrößte Gruppe bilden die *prüfungsbezogenen Prädikate*. Hier lässt sich größtenteils eine Zuordnung zu typischen Aufgabenstellungen treffen, die im Rahmen der Auslegung auftreten:

(a) Kopula-ähnliche Verben, die wie *gelten* oder *ansehen als* im Nebensinn eine propositionale Einstellung zum Ausdruck bringen, kennzeichnen oft Angaben von Bedingungen, unter denen vom etablierten, insbesondere vom alltagssprachlich bekannten Sinn des Definiendum abgewichen werden kann (z.B. sogenannte juristische Fiktionen oder Definitionen zur Regelung von Normkonkurrenzen, vgl. 1.1.2).

(b) Prädikate wie *Voraussetzung ist* oder *ausreichend ist* thematisieren den *logischen Status* des Definiens und werden verwendet, wenn die argumentative Absicherung einer Entscheidung Hauptfunktion einer Definition ist.

(c) Die dritte Hauptgruppe prüfungsbezogener Prädikate leitet *prozedurale Angaben* oder *Angaben von Prüfungsmaßstäben* ein, die vor allem im Zusammenhang mit evaluativen oder intentionalen Konzepten wichtig sind.

Die übrigen in Tabelle 2.4 genannten prüfungsbezogenen Prädikate dienen zu spezifischeren Zwecken, nämlich zur Formulierung (d) fallbezogener An-

gaben zu Begriffsumfang bzw. Extension des Definiendum und (e) expliziter Angaben zu Begriffsbeziehungen (meist zwischen institutionellen Begriffen).

Bei den *semantisch leichten* und den *Merkmal-spezifischen* Prädikaten lassen sich breit anwendbare Prädikate von solchen unterscheiden, die nur mit spezifischen Definienda auftreten. In der ersten Gruppe sind zum Beispiel die Verben *vorliegen* (das nach Auftreten zweithäufigste Definitionsverb im untersuchten Korpus) und *gegeben sein* mit einer Vielzahl nominaler Definienda kombinierbar, während das Verb *genießen* mit dem Nomen *Recht* und dessen Komposita kollokativ assoziiert ist. In der zweiten Gruppe treten neben generischen Kategorieangaben (z.B. *Bestandteil* bei dem Prädikat *gehören (zu)*) auch Angaben spezifischer Merkmalstypen auf, die nur aufgrund fachspezifischer Zusammenhänge für ganz bestimmte Begriffe definitionsrelevant sind (z.B. wird mit dem Verb *schützen* der Schutzbereich oder -zweck bei der finalen Bestimmung einiger Grundrechte angegeben).

Die starke semantische Ausdifferenzierung innerhalb der Gruppe der prüfungsbezogenen Prädikate erklärt sich, wenn man die Verteilung der unterschiedenen Prädikatklassen auf die in 2.2 diskutierten Schichten *Kerndefinition* und *Elaboration* betrachtet. Diese ist für die von uns untersuchten Dokumente in Abb. 2.2 dargestellt.

Explizit bedeutungsbezogene und Klassifikationsprädikate werden hauptsächlich zur Formulierung von Kerndefinitionen verwendet. Definitionen mit prüfungsbezogenen Prädikaten machen dagegen den weitaus größten Teil der elaborierenden Angaben aus. Definitionen dieser Schicht erfüllen – wie in 2.2 ausführlicher diskutiert – eine große Zahl verschiedener Aufgaben und erfordern daher auch auf der Formulierungsebene entsprechende Differenzierungsmöglichkeiten.

#### (d) Realisierung von Definitionsbestandteilen

Während bei appositiven Definitionen keine allgemeinen Regeln zur Lokalisierung von Definiendum und Definiens angegeben werden können, korrespondieren die Definitionsbestandteile bei prädikatbasierten Definitionen typischerweise zu dem Definitionsprädikat untergeordneten syntaktischen Einheiten innerhalb der Definition.

**Definiendum.** Bei den oben genannten Prädikaten entspricht das Definiendum durchgängig einem semantischen Argument des Definitionsprädikats (welchem, hängt von dessen Semantik ab). Wie es realisiert werden kann ist also durch das mit dem Prädikat verknüpfte Linking-Muster und den Valenzrahmen vorgegeben. So ist das Definiendum z.B. in Definitionen mit dem Verb *vorliegen* in der (meistens indefiniten) Subjekt-Nominalphrase enthalten, während es

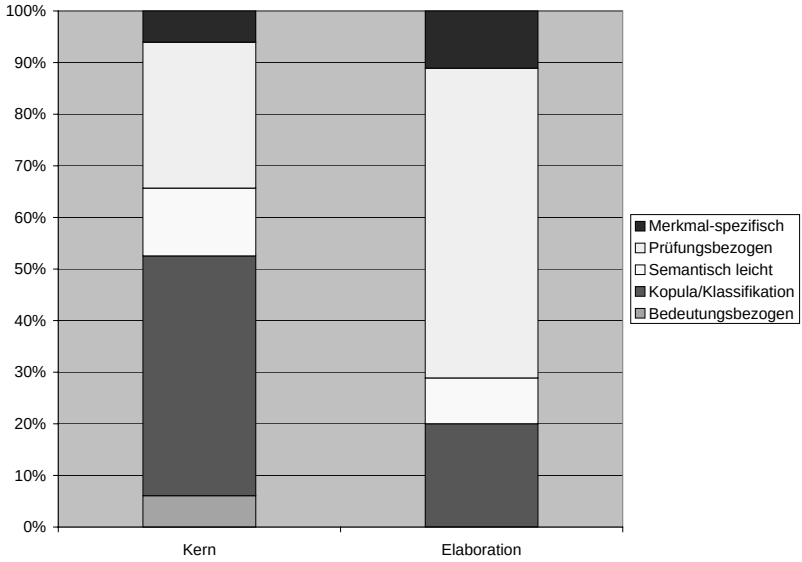



Abbildung 2.2: Definitionsprädikate in Kerndefinitionen und elaborierenden Aussagen

dem Verb *verstehen* als subkategorisierte Präpositionalphrase mit der Präposition *unter* untergeordnet wird, da das Subjekt hier eine agentive und intentionale Entität sein muss. In Tabelle 2.5 ist angegeben, welche Satzteile bei den Definitionsprädikaten aus Tabelle 2.4 in unserem Korpus das Definiendum enthalten.

In der als Definiendum fungierenden Phrase kann oberhalb oder unterhalb des eigentlichen Definiendum weiteres Material integriert sein. Oberhalb des Definiendum finden sich semantisch weitgehend transparente Substantive (wie *Begriff*, vgl. (2.16)) oder solche Substantive, die die Funktion des Definiendum hinsichtlich einer Norm oder einer anderen Definition beschreiben (z.B. *Tatbestandsmerkmal* in (2.17)). Dem Definiendum untergeordnet stehen nichtrestriktive Relativsätze und Quellenangaben (meist – wie in (2.1) – eingeleitet durch die Formel *im Sinne des/von*). Außerdem kann das Definiendum in einen Anwendungskontext gestellt werden, der zugleich die Geltung der Definition auf einen bestimmten Sachbereich einschränkt. So wird in (2.18) das

| **Definiendum** | **Prädikate** |
|---|---|
| Subjekt | *ausschließen, bedeuten, besagen, der Fall sein, entfallen, erfordern, folgen, fordern, Gegenstand sein, müssen, schützen, sich begnügen, sich beziehen, sich richten, sicherstellen, umfassen, umschreiben, verlangen, voraussetzen, Voraussetzung sein, vorliegen, zukommen* |
| Akkusativobjekt | *anerkennen, annehmen, ausschließen, beurteilen, darstellen, entnehmen, ermitteln, geben, genießen, haben, kennzeichnen* |
| Dativobjekt | *genügen* |
| Präpositionalphrase | *ausgehen, Ausnahmen geben, ausreichend sein, einbeziehen, entscheidend sein, erforderlich sein, gehören, genügen, gleichgültig sein, maßgebend sein, sich handeln, sprechen, verstehen, Voraussetzung sein, werden, zählen* |
| *als*-Prädikativum | *ansehen, betrachten, gelten, verstehen* |
| Prädikatsnomen | *werden* |
| Infinitivsatz | *müssen* |

Tabelle 2.5: Realisierung des Definiendum

Adjektiv *unwesentlich* nur in seiner Anwendung auf den Begriff Beeinträchtigung bestimmt, dessen Bedeutung nicht “mitdefiniert”, sondern als bekannt vorausgesetzt wird.

(2.16) Der Begriff “Schichtarbeit” liegt nicht nur dann vor, wenn ein Arbeitnehmer das begonnene Arbeitsergebnis des anderen mit denselben Mitteln oder der gleichen Intensität und Belastung vervollständigt, sondern auch dann, wenn ein gewisses Maß an Arbeitsteilung für ein und denselben Arbeitserfolg erforderlich ist, die verschiede-

nen Arbeitsergebnisse also aufeinander aufbauen (BAG Urteil vom 20. Dezember 1961, aaO).

(BAG 6. Senat, 4. Februar 1988, AZ 6 AZR 203/85, juris)

(2.17) Wie oben ausgeführt, liegt das Tatbestandsmerkmal, daß eine nachteilige Veränderung des Grundwassers "nicht zu besorgen ist", nicht schon dann vor, wenn eine Gefährdung aller Voraussicht nach nicht eintreten wird.

(Verwaltungsgerichtshof Baden-Württemberg 3. Senat, 3. August 1998, AZ 3 S 990/98, juris)

(2.18) Danach liegt eine nur unwesentliche Beeinträchtigung in der Regel dann vor, wenn Immissionen die in Gesetzen oder Rechtsverordnungen festgelegten Grenz- oder Richtwerte nicht überschreiten.

(Hamburgisches Oberverwaltungsgericht 3. Senat, 20. Januar 1997, AZ Bf III 54/95 P, juris)

Bei den meisten der oben genannten Definitionsprädikate erfordert die als Definiendum fungierende Valenzstelle einen nominalen Ausdruck, so dass nur nominale Definienda direkt definiert werden können (die wichtigsten Ausnahmen sind die Kopula *sein* und die Prädikate mit prädikativem Definiendum, die auch Definitionen adjektivischer Ausdrücke erlauben). Um nicht-nominale Termini mittels dieser Definitionsprädikate zu bestimmen, müssen die Termini (regulär wie in (2.11) oder durch Derivation wie in (2.3)) nominalisiert, oder (wie in (2.17) und (2.18)) unter ein transparentes Nomen eingebettet bzw. mit einem Nomen als Bezugswort kontextualisiert werden. Vereinzelt finden sich außerdem verkürzte Formulierungen am Rande der Grammatikalität, in denen z.B. wie in (2.19) nicht-nominale Termini an Komplementstellen auftreten, die eigentlich nur von nominalen Ausdrücken gefüllt werden können.

(2.19) Eigenhändig bedeutet, dass der Aussteller bzw. Urheber der Erklärung [. . . ] selbst mit der Hand die Namensunterschrift leisten muss.

(Landesarbeitsgericht Köln 13. Kammer, 19. Juni 2001, AZ 13 Sa 1571/00, juris)

**Definiens.** Auch die Realisierung des Definiens hängt sowohl von der Semantik des Definitionsprädikats ab als auch davon, welche Ergänzungen und Angaben dieses lizensiert. Die sich aus der Bedeutung des jeweiligen Prädikats ergebende Definitionsmethode bestimmt darüber, welchen semantischen Typ das Definiens haben muss. Das Verb *vorliegen* thematisiert beispielsweise

direkt die Frage der Anwendbarkeit oder Nicht-Anwendbarkeit eines Begriffes. Mit diesem Verb werden daher (wie auch mit vielen der oben genannten prüfungsbezogenen Prädikate) Definitionen durch Angabe von *Anwendungsbedingungen* getroffen. Im Gegensatz dazu wird zum Beispiel mit der Kopula *sein* oder dem Prädikat *verstehen (unter)* durch *Klassifikation* und mit dem Verb *bedeuten* durch *Umschreibung* des Definiendum definiert. Bei Definitionen mit *vorliegen* muss das Definiens daher auf eine vollständige Aussage über das Definiendum zurückgeführt werden können. Dagegen muss es in Definitionen mit *sein* bzw. *verstehen (unter)* über das Definiendum prädizierbar sein und bei *bedeuten* dem semantischen Typ des Definiendum entsprechen. Darüber, wie ein konkreter Inhalt so realisiert werden kann, dass die Konstruktion den jeweiligen semantischen Anforderungen entspricht, bestimmt wiederum der Valenzrahmen des jeweiligen Definitionsprädikats. So werden z.B. auch mit *erfordern* Definitionen auf der Basis einer Anwendungsbedingung formuliert. Mit *erfordern* wird diese jedoch im Objekt angegeben, während sie mit dem Verb *vorliegen*, das nicht für ein Objekt subkategorisiert, als freie Angabe kombiniert wird.

Definitionsmethode und Valenzrahmen des Definitionsprädikats determinieren die möglichen Formulierungen für das Definiens allerdings vielfach nicht vollständig. So sind sowohl bei *erfordern* als auch bei *vorliegen* jeweils eine satzförmige (*erfordern, dass* bzw. *vorliegen, wenn*) und eine nominale Realisierung (*erfordern + Akkusativ* bzw. *vorliegen bei*) der Anwendungsbedingung möglich. Für *vorliegen* sind die beiden Varianten in (2.20) bzw. (2.21) belegt:

(2.20) Dagegen liegt eine Enteignung dann vor, wenn der belastende Rechtsakt eine duch [sic] Art. 14 Abs. 1 Satz 1 GG geschützte Rechtsposition ganz oder teilweise entzieht.

(VG Meiningen 2. Kammer, 13. Juli 1994, AZ 2 K 18/93.Me, juris)

(2.21) Eine Enteignung im eigentlichen Sinne liegt nur bei hoheitlichen Akten vor, die darauf gerichtet sind, dem einzelnen konkrete, als Eigentum geschützte Rechtspositionen zur Erfüllung bestimmter öffentlicher Aufgaben vollständig oder teilweise zu entziehen (. . . )

(BGH 3. Zivilsenat, 20. Januar 2000, AZ III ZR 110/99, juris)

Die beiden Beispiele lassen erkennen, dass zwischen der nominalen und klausalen Formulierungsvariante (zumindest in Definitionen mit dem Verb *vorliegen*) kein relevanter semantischer Unterschied besteht – (2.20) und (2.21) liefern weitgehend gleichwertige Definitionen des Terminus *Enteignung*. Allerdings enthält das Definiens in (2.21) verglichen mit (2.20) ausführlichere

Angaben, und es ist zu vermuten, dass für diese aus stilistischen Gründen eine klausale Realisierung gewählt wurde.

Tabelle 2.6 gibt für die Prädikate aus Tabelle 2.4 die syntaktischen Positionen an, in denen in den von uns untersuchten Definitionen das Definiens auftrat.

| **Definiens** | **Prädikate** |
|---|---|
| Subjekt | *ausreichend sein, ausschließen, darstellen, gehören, gelten, genießen, maßgebend sein, werden, zählen* |
| Akkusativobjekt | *ansehen, ausschließen, betrachten, einbeziehen, erfordern, fordern, schützen, sicherstellen, umfassen, umschreiben, verlangen, verstehen, voraussetzen* |
| Dativobjekt | *entnehmen* |
| Präpositionalphrase | *beurteilen, ermitteln, folgen, kennzeichnen, sich begnügen, sich beziehen, sich handeln, sich richten, vorliegen* |
| Konditionalsatz | *anerkennen, annehmen, ansehen, ausgehen, Ausnahmen geben von, der Fall sein, entfallen, geben, genügen, haben, sprechen, vorliegen, werden, zukommen,* |
| dass / ob-Satz | *ausreichend sein, bedeuten, besagen, entscheidend sein, erforderlich sein, genügen, gleichgültig sein, Voraussetzung sein, voraussetzen* |
| Hauptsatz | *müssen* |
| Infinitivsatz | *Gegenstand sein* |

Tabelle 2.6: Realisierung des Definiens

**Zusatzinformation.** Außer Definiendum, Definitionsprädikat und Definiens können juristische Definitionen noch weitere systematisch relevante Informationseinheiten enthalten. In 2.2 sind wir bereits auf die Funktion und Realisierungsmöglichkeiten von *Angaben zum argumentativen Status* und *Angaben zum Geltungsbereich* eingegangen. Neben den dort besprochenen Sachbereichsangaben kann der Geltungsbereich einer Definition auch durch Rechtsbereichs-

und Quellenangaben bestimmt werden. Wie Sachbereichsangaben können diese mit syntaktischen Mitteln (meist als Präpositionalphrase mit der Präposition *nach* oder mittels der bereits oben erwähnten Formel *im Sinne des / von*) in die Definition integriert werden, wobei zwischen Angaben zum definierten Terminus und Angaben zur ganzen Definition unterschieden werden kann.

Weitere wichtige Bestandteile vieler Definitionen sind:

1. Taxonomische Information
2. Diskursstrukturierende Elemente
3. Geltungsregulierende Elemente

*Taxonomische Information* stellt zum Beispiel der Genusbegriff in Definitionen nach dem *genus et differentia*-Schema dar. Allerdings liefern einerseits (wie bereits diskutiert) nicht alle Definitionen auf der Basis der Kopula *sein* einen taxonomisch sinnvollen Genusbegriff, anderseits kann taxonomisch verwertbare Information auch in anderer Form aus Definitionen zu entnehmen sein. So enthalten (2.14) (in zwei Definitionen) und (2.16) (innerhalb des Definiendum) Bestimmungen zweier einander ausschließender gleichgeordneter Konzepte mit einem gemeinsamen Oberbegriff.

In (2.16) wird der Alternativen-Charakter der beiden Bestimmungen zudem durch die kontrastiven *diskursstrukturierenden* Elemente *nicht nur. . . sondern auch* markiert, die hier korrelativ in einem Satzgefüge stehen. Meistens haben solche Ausdrücke jedoch eine satzexterne Funktion und dienen somit gleichzeitig der kohärenten Eingliederung der jeweiligen Definition in ihren Kontext. So wird in (2.20) durch das Diskurskonnektiv *dagegen* ein Kontrast zu einer vorangehenden Ausschlussbedingung zum selben Terminus hergestellt und die Definition damit in den Argumentationsverlauf eingeordnet. Bei Ausdrücken wie *insoweit* (vgl. (2.4)) und *dabei* (vgl. (2.9)) überwiegt diese Kohärenzherstellungsfunktion, und sie sind für die eigentliche Definition inhaltlich ohne Bedeutung.

*Geltungsregulierende Elemente* sind Ausdrücke wie *in der Regel* in (2.18). In der linguistischen Literatur werden solche (in Anlehnung an Lakoff (1973) als *Hedges* bezeichneten) Elemente vor allem im Hinblick auf ihre sprecher- und adressatenbezogene Funktion analysiert. Wie im vorigen Kapitel erläutert wirken Definitionen in Urteilsbegründungen vielfach in der nachfolgenden Rechtsprechung fort und sind (außer bei höchstrichterlichen Entscheidungen) jedenfalls potentiell der Revision durch nachfolgende Instanzen unterworfen. Die *Hedging*-Funktion geltungsregulierender Elemente besteht hier daher vorwiegend darin, dass der Richter die “juristischen Nachwirkungen” seiner Äußerun-

gen kontrolliert, um sich damit vor den institutionellen Folgen möglicherweise nicht haltbarer Konsequenzen aus seinen Festlegungen zu schützen.[7]

### 2.3.3 Fazit

Definitionen in Urteilsbegründungen lassen sich nach der Art der sprachlichen Verknüpfung in appositive bzw. parenthetische und in prädikatbasierte Definitionen einteilen. Während appositiv bzw. parenthetisch realisierte Definitionen oft kommentarartig zur nochmaligen Anführung bekannten terminologischen Wissens verwendet werden, werden mit prädikatbasierten Definitionen voll in den Textzusammenhang eingebundene terminologische Feststellungen, Präzisierungen oder auch Neufestlegungen als Hauptaussage eines Satzes getroffen.

Die Variationsbreite der verschiedenen in solchen Definitionen verwendeten Prädikate ist in Urteilsbegründungen deutlich größer als dies in bisherigen Untersuchungen für technischen und wissenschaftlichen Text festgestellt wurde. Eine Systematisierung nach semantisch-funktionalen Gesichtspunkten lässt erkennen, dass Prädikate mit Bezug zum juristischen Prüfungs- und Argumentationsverlauf eine stark ausdifferenzierte Gruppe von Definitionsprädikaten darstellen, die für die Elaboration juristischer Begrifflichkeiten eine besonders wichtige Rolle spielt.

Die Definitionsbestandteile korrespondieren in prädikatbasierten Definitionen in der Regel zu einzelnen Satzteilen. Ihre konkrete Realisierung resultiert allerdings stets aus dem Zusammenspiel verschiedener Faktoren. Es liegen deshalb keine generellen (etwa semantischen) Kriterien für eine übergreifende Systematisierung der Varianten dieser Abbildung auf der Hand. Eingeschränkt auf Gruppen bestimmter Definitionsprädikate lassen unsere Beispiele hingegen durchaus klare Regelmäßigkeiten bei der Realisierung für Definiendum und Definiens erkennen (im Gegensatz zu den im vorigen Abschnitt zuletzt besprochenen Zusatzangaben). Oft ist das gemeinsame Auftreten eines Definitionsprädikats mit den entsprechenden Ergänzungen bzw. Angaben dabei zugleich ein guter Indikator für das Vorliegen einer Definition. In Kap. 5 befassen wir uns mit der Nutzung solcher Kombinationen als Suchmuster in einem System zur automatischen Identifikation von Definitionen in Urteilstexten.

[7] Allerdings wird die Bindungswirkung von Begriffsbestimmungen außer durch explizite Geltungsregulierung auch durch die Autorität der Quelle beeinflusst. Wird eine Begriffsverwendung vom Bundesgerichtshof als *in der Regel* geltend gekennzeichnet, kann dies unter Umständen für untergeordnete Gerichte eine voll verbindliche Festlegung sein, während auch eine von einem unteren Gericht als vollgültig vorgetragene Definition durch höhere Gerichte revisibel bleibt.

## 2.4 Annotationsstudie

In den vorigen Abschnitten haben wir anhand von Beispielen einen verglichen mit der klassischen Definitionslehre erweiterten juristischen Definitionsbegriff charakterisiert. Um die bei der Anwendung dieses Begriffs erreichbare Stabilität und intersubjektive Verlässlichkeit einschätzen zu können, haben wir nach der Auswertung der in 2.1 beschriebenen Daten eine zweite Annotationsstudie mit zwei Annotatoren auf einer breiteren Datengrundlage durchgeführt. Wir haben so zugleich einen Goldstandard für die Evaluation automatischer Extraktionstechniken erzeugt. Im Rest dieses Kapitels gehen wir auf die Durchführung und die Ergebnisse dieser Studie ein.

### 2.4.1 Annotationsszenario und -richtlinien

Als Datengrundlage für diese Untersuchung diente eine Auswahl von sechzig Urteilsbegründungen (im Folgenden auch: *Goldstandard-Korpus*) aus der Entscheidungssammlung der Firma *juris*, wiederum verteilt über verschiedene Gerichtsbarkeiten und Sachgebiete. Es wurden nur Urteile mit vollständig erfassten Entscheidungsgründen in Betracht gezogen, der Schwerpunkt lag auf Urteilen von Bundesgerichten. Die Verteilung der ausgewählten Urteile auf Gerichtsbarkeiten und Sachgebietskategorien ist Tabelle 2.7 zu entnehmen.

| **Gerichtsbarkeit** | **Anzahl** | **Sachgebiet** | **Anzahl** |
|---|---|---|---|
| BVerfG | 14 | Arbeitsrecht | 7 |
| BGH | 16 | Allgemeines Verwaltungsrecht | 25 |
| BFH | 7 | Bürgerliches Recht | 10 |
| BSG | 6 | Besonderes Verwaltungsrecht | 46 |
| BAG | 3 | Weitere Gerichtsverfahrensordnungen | 13 |
| LVerfG | 1 | Gerichtsverfassung und -zuständigkeit | 1 |
| OLG | 1 | Handels- und Wirtschaftsrecht | 9 |
| VGH | 4 | Regionen, Rechtswissenschaft, | 23 |
| OVG | 3 | Kirchenrecht | |
| LArbG | 3 | Sozialrecht | 11 |
| LSG | 2 | Strafrecht | 4 |
| | | Steuerrecht (einschl. Steuerberatungsrecht) | 8 |
| | | Staats- und Verfassungsrecht | 23 |
| | | Zivilprozessrecht | 6 |
| | | Recht des öffentlichen Dienstes | 6 |

Tabelle 2.7: Zuordnung der Entscheidungen im Goldstandard-Korpus zu Gerichtsbarkeiten und Sachgebietskategorien

Die ausgewählten Entscheidungen umfassten insgesamt 275 901 Textwörter in 9421 Sätzen, nach Entfernen der nicht begründungsrelevanten Textteile verblieben 7627 Sätze mit 233 210 Textwörtern. In diesem Korpus wurden unter Verwendung des bereits in 2.1 beschriebenen Annotationsschemas definitorische Textpassagen gekennzeichnet, und es wurden in den annotierten Textspannen auf Satzebene Kerninformation und elaborierende Angaben separat markiert. Die Kategorien für die Einordnung der argumentativen Funktion von Definitionen (vgl. 2.2.1) haben wir erst nach Abschluss der hier beschriebenen Studie erarbeitet. Diese Information wurde daher im Goldstandard-Korpus nicht annotiert.

Als Entscheidungsgrundlage bei der Annotation diente im Unterschied zu der in 2.1 diskutierten explorativen Studie ein schriftliches Annotationshandbuch, Walter (2006). Dieses gab die in 2.2.4 angeführten Kriterien wieder und erläuterte sie. Ergänzend waren kommentierte Positiv- und Negativbeispiele angegeben, die die einzelnen Kriterien verdeutlichen und Entscheidungshilfe bei typischen Abgrenzungsproblemen zwischen den Informationstypen *Kerndefinition* und *Elaboration* sowie den drei möglichen Konfidenzstufen *klar, unklar* und *zweifelhaft* bieten sollten.

Die Annotation des Korpus wurde unabhängig voneinander von einem juristischen Annotator (Jurist nach dem ersten Staatsexamen, im Folgenden *J*) und – aus linguistischer Perspektive – vom Autor dieser Arbeit (im Folgenden *L*) vorgenommen. Nach Abschluss der doppelten Annotation wurden die Ergebnisse von beiden Annotatoren gemeinsam gesichtet und zu einem Goldstandard zusammengeführt. Dabei bestand für sämtliche annotierte Information, auch bei Übereinstimmung der ursprüngliche Annotationen, die Möglichkeit der Revision. Im Rahmen dieser Zusammenführung wurde außerdem für die einzelnen Definitionen bestimmt, ob sie dem appositiv-parenthetischen oder dem prädikatbasierten Realisierungstyp zuzuordnen sind.

### 2.4.2 Ergebnisse

Eine Übersicht über die Ergebnisse dieser Annotationsstudie gibt Tabelle 2.8.

Wie schon bei der Annotation des Pilotstudien-Korpus führte die Beschränkung auf Sätze als kleinste annotierbare Einheiten nicht zu nennenswerten Problemen. Der Anteil diskontinuierlicher Definitionskomplexe (J: 5; L: 7; Goldstandard: 6) und der Anteil der Sätze, die mehreren Definitionskomplexen angehören (J: 11, L: 4, Goldstandard: 14) war geringer als dort. Das gleiche gilt für den Anteil von Sätzen mit mehr als einer Definition (J: 10, L: 4, G: 13). Wie schon für das Pilotstudien-Korpus haben wir auch hier bei der Auswertung in diesen Fällen nur die jeweils prominenteste Definition im Satz betrachtet.

| | *Annotator J* | *Annotator L* | *Goldstandard* |
|---|---|---|---|
| *Anzahl* | 261 | 317 | 274 |
| *klar / unklar / zweifelhaft* | 99 / 98 / 64 | 154 / 91 / 72 | 177 / 74 / 23 |
| *∅ Sätze / Def.* | 1,9 | 1,8 | 1,9 |
| *1 Satz* | 57% | 62% | 59% |
| *2 Sätze* | 21% | 17% | 17% |
| *3 Sätze* | 11% | 10% | 12% |
| *>3 Sätze* | 10% | 11% | 12% |
| *Gesamtzahl Sätze* | 483 | 571 | 507 |

Tabelle 2.8: Übersicht über die Annotationsergebnisse

Von den 507 in den Goldstandard übernommenen Sätzen wurden 473 als prädikatbasierte Definitionen identifiziert.[8]

Sowohl die Rate an Definitionen pro Dokument (J: 4,4; L: 5,3; Goldstandard: 4,6) als auch der Anteil der als definitorisch markierten Sätze am Gesamtkorpus (J: 6,3%; L: 7,5%; Goldstandard: 6,7%) lag deutlich höher als in der vorangegangenen Studie (3,45 Definitionen / Dokument, 5% der Sätze definitorisch).[9] Die Abweichung im Hinblick auf die Anzahl der Definitionen pro Dokument erklärt sich dadurch, dass ein Teil der Dokumente im Pilotstudien-Korpus bei *juris* nur in einer Kurzfassung erfasst ist und daher nicht der volle Entscheidungstext zur Annotation vorlag. Der im Vergleich zur ersten Annotation deutlich höhere Anteil definitorischer Sätze am Gesamtkorpus geht dagegen darauf zurück, dass beide Annotatoren in der zweiten Studie generell längere Textpassagen als Definitionen kennzeichneten als der Annotator in der ersten Studie (wo die durchschnittliche Länge einer Definition 1,4 Sätze betrug).

Wie der Tabelle 2.9 entnommen werden kann, wurde vor allem eine im Verhältnis zur definitorischen Kerninformation deutlich größere Zahl elaborieren-

[8] Die nicht prädikatbasierten Formulierungen traten hier wie im Pilotstudien-Korpus fast ausschließlich in einsätzigen Definitionen auf. Berücksichtigt man nur diese als Grundgesamtheit, so liegt der Anteil appositiv-parenthetischer Definitionen in beiden Korpora in einem ähnlichen Bereich.

[9] Für prädikatbasierte Definitionen ergibt sich unter Absehung von Sätzen die als *zweifelhaft* annotiert sind und bei Einschränkung der Grundgesamtheit auf begründungsrelevante Sätze in beiden Korpora zusammen eine mittlere "Definitionsquote" von 3,6%. Von dieser gehen wir in Kap. 5 als Erwartungswert für Recall-Schätzungen bei der Definitionsextraktion aus.

| | *Annotator J* | *Annotator L* | *Goldstandard* |
|---|---|---|---|
| *Kern* | 263 | 217 | 242 |
| *Anteil am Gesamtkorpus* | 3,4% | 2,8% | 3,1% |
| *Elaboration* | 220 | 354 | 265 |
| *Anteil am Gesamtkorpus* | 2,9% | 4,6% | 3,5% |

Tabelle 2.9: Anteil von Kern- und elaborierenden Aussagen

der Aussagen annotiert. Während der Anteil definitorischer Kernaussagen am Gesamtkorpus für beide Annotatoren und den Goldstandard ähnlich groß ist wie in der ersten Studie (3,4%), liegt der Anteil der elaborierenden Aussagen jeweils deutlich über den dort ermittelten 1,6%.

#### (a) Übereinstimmung der Annotatoren

Schon aus Tabelle 2.8 und Tabelle 2.9 wird deutlich, dass Annotator J zwar insgesamt strengere Anforderungen für die Klassifikation einer Textspanne als Definition stellte als Annotator L, letzterer jedoch für die Auszeichnung als Kerninformation deutlich strengere Kriterien ansetzte. Der gemeinsam erstellte Goldstandard nimmt im Hinblick auf alle bisher betrachteten Kriterien eine Mittelstellung zwischen Annotation J und L ein.

Eine genauere quantitative Bewertung der Übereinstimmung der Annotatoren ist aufgrund der multidimensionalen und auf Textspannen bezogenen Annotation nicht direkt durch einen einzelnen Koeffizienten möglich, sondern erfordert eine mehrstufige Vorgehensweise. Wir betrachten im Folgenden zunächst auf Satzebene den Grad der Übereinstimmung der Annotationen bei der binären Entscheidung *definitorisch / nicht definitorisch*, dann den Grad der Übereinstimmung bei der Bewertung der überlappenden Sätze als Kerninformation bzw. elaborierende Aussage und schließlich den Überlappungsgrad der als Definitionen gekennzeichneten Textspannen.

Die Auswertung der Übereinstimmung bei der Klassifikation der 7626 Sätze des annotierten Korpus ergab die in Tabelle 2.10 aufgeführten κ-Werte.[10]

[10] Der in Cohen (1960) zur Bewertung der Übereinstimmung zweier Rater vorgeschlagene Koeffizient κ berechnet sich nach der Formel $\frac{p_0 - p_c}{1 - p_c}$, wobei $p_0$ die beobachtete und $p_c$ die zufällig zu erwartende Übereinstimmung bezeichnet. Letztere wird anhand der Gesamtverteilung des beurteilten Merkmals in den Bewertungen abgeschätzt.

| | κ |
|---|---|
| [±*definitorisch*] | 0,58 |
| *Kern / Elaboration* | 0,56 |

Tabelle 2.10: Übereinstimmung der Annotatoren J und L

Die Übereinstimmung der Annotatoren bei der Entscheidung *definitorisch / nicht definitorisch* liegt am oberen Rand des von Landis und Koch (1977) als *moderate* und von Greve und Wentura (1997) als *akzeptable* Übereinstimmung eingestuften Bereichs. Die Übereinstimmung mit dem erzeugten Goldstandard liegt für beide Annotatoren in einem ähnlichen Bereich (0,82 für L bzw. 0,86 für J), was darauf hindeutet, dass dieser eine relativ genaue Mittelstellung zwischen den unterschiedlichen Annotationen einnimmt.

Bezogen auf einzelne Dokumente ergeben sich für die Korrelation der beiden Annotationen hinsichtlich des Merkmals [±definitorisch] κ-Werte von 0 (entsprechend der zufällig zu erwartenden Übereinstimmung) bis 1 (volle Übereinstimmung), wobei der Median der Verteilung bei 0,63 liegt. Jedenfalls für die Hälfte der Dokumente wird also ein Übereinstimmungsgrad erzielt, der laut Landis und Koch (1977) als *substantiell* einzustufen ist. Läßt man das Fünftel der Dokumente mit besonders geringer Übereinstimmung zwischen den Annotatoren (entsprechend einem Schwellenwert von κ=0,3) außer Betracht, ergibt sich auch als Gesamtwert eine *substantielle* Übereinstimmung von κ=0,64.

Eine systematische Einschätzung der chronologischen Entwicklung des Übereinstimmungsgrades ist nicht möglich, da die zu annotierenden Dokumente von den Annotatoren in selbst gewählter Reihenfolge bearbeitet werden konnten. Der Vergleich der Übereinstimmung auf fünf von beiden Annotatoren jeweils im ersten bzw. letzten Drittel der Annotationsphase bearbeiteten Dokumenten (κ=0,48 bzw. κ=0,6) deutet jedoch auf einen gewissen Trainings- und Konsolidierungseffekt bei der Klassifikation hin.

Auch bei der Klassifikation definitorischer Sätze als Kerninformation bzw. elaborierende Aussage liegt der Grad der Übereinstimmung der Annotatoren J und L (bezogen auf die überhaupt von beiden als definitorisch eingestuften Sätze) am oberen Rand des als *moderat* bzw. *akzeptabel* zu bewertenden Bereichs.

| | *J-L* | *L-J* |
|---|---|---|
| *Überlappung* | 69% | 59% |

Tabelle 2.11: Gesamtüberlappung aller annotierten Definitionen

Betrachtet man für beide Annotatoren den Anteil der als Definition markierten Textspannen, die sich mit einer vom jeweils anderen Annotator gekennzeichneten Definition überschneiden (Tabelle 2.11), ergibt sich für Annotator J ein deutlich höherer Wert als für Annotator L. Dies spiegelt dessen (wie bereits angesprochen) "liberalere" Annotationspraxis wider. Gleiches gilt für den Anteil der Definitionen aus Annotation J bzw. L, die sich mit in den Goldstandard übernommenen Textspannen überschneiden. Legt man jeweils für einen der beiden Annotatoren nur die als *Kern* klassifizierten Sätze bei der Bestimmung der Überschneidungen zu Grunde, gleichen die offenbar strengeren Kriterien von Annotator L bei der Einstufung als *Kerninformation* die angesprochene Tendenz jedoch wieder aus: Der Anteil der überlappenden Definitionen an den insgesamt von J identifizierten liegt nach dieser Auswertungsmethode mit 67% deutlich niedriger als für L mit 75%. In den sich überschneidenden Definitionen selber ist in allen Fällen die Überlappung größtenteils vollständig, der durchschnittliche Überschneidungsgrad liegt für alle Konstellationen bei über 90%.

#### (b) Vergleich mit dem *juris*-Definitionsregister

Die Firma *juris* führt für einen Teil der dokumentierten Entscheidungstexte ein Definitionsregister. Dabei handelt es sich um ein Schlagwortverzeichnis, in dem diejenigen Begriffe vermerkt werden, zu denen im Text einer katalogisierten Entscheidung nach Einschätzung des bearbeitenden Dokumentars definitorische Angaben gemacht werden. Der bei der Erstellung des Registers zu Grunde gelegte Definitionsbegriff ist allerdings nicht dokumentiert, so dass die Möglichkeit eines Vergleichs mit der hier diskutierten Annotationsstudie auf konzeptueller Ebene (etwa durch Gegenüberstellung von Annotationsrichtlinien) ausscheidet.

Für alle sechzig von uns zur Annotation ausgewählten Entscheidungstexte sind Einträge im *juris*-Definitionsregister vorhanden. Auch auf der Ebene der für die einzelnen Dokumente ermittelten Definitionen ist jedoch kein direkter

Vergleich unserer Annotationsergebnisse mit dem Definitionsregister möglich, da Entscheidungen von *juris* auf Dokumentebene indiziert werden, d.h. keine Zuordnung von Indextermen zu bestimmten Textpassagen oder einzelnen Sätzen erfolgt. Wir haben im Anschluss an die Zusammenführung der Annotationen J und L den in den Goldstandard übernommenen Definitionen manuell soweit wie möglich die entsprechenden bei *juris* für das jeweilige Dokument verzeichneten Indexterme zugeordnet. Die auf der Basis dieser Zuordnung ermittelten Daten zur Übereinstimmung unserer Annotation mit dem bei *juris* geführten Register sind in Tabelle 2.12 zusammengefasst:

| | *Anzahl* | *Anteil* |
|---|---|---|
| *Indexeinträge* | 91 | |
| *davon im Goldstandard definiert (≙ Recall)* | 84 | 92% |
| *Definitionen für Indexterme im Goldstandard (≙ Präzision)* | 136 | 49% |
| *Definitionen je Indexeintrag* | 1,6 | |

Tabelle 2.12: Übereinstimmung der Annotation mit dem *juris*-Definitionsregister

Für fast alle im *juris*-Definitionsregister verzeichneten Termini wurden demnach in unserer Annotationsstudie tatsächlich Definitionen identifiziert. Dies deutet darauf hin, dass der Erstellung des *juris*-Definitionsregisters implizit ähnliche Kriterien zu Grunde liegen wie unserer Annotationsstudie. Der hohe Anteil annotierter Definitionen ohne entsprechende Indexeintrag legt andererseits den Schluss nahe, dass der durch das Definitionsregister reflektierte Definitionsbegriff deutlich enger gefasst ist als der hier erarbeitete. Allerdings ist die Praxis bei der Pflege des *juris*-Definitionsregisters offenbar so uneinheitlich, dass nur schwer einzuschätzen ist, ob es überhaupt sinnvoll als Referenz für einen Vergleich wie den hier diskutierten herangezogen werden kann. Zum einen erfolgen Eintragungen in das Definitionsregister nach Auskunft von *juris* generell fakultativ, so dass nicht angenommen werden kann, dass überhaupt alle in einer Entscheidung definierten Begriffe zwangsläufig in das Register übernommen werden. Die Untersuchung einzelner Einträge zeigt zum anderen, dass sich das Register in vielen Fällen auf die in einem Urteilstext zitierten Legaldefinitionen beschränkt. Andererseits werden aber auch immer wieder andere fallrelevante Rechtsbegriffe aufgenommen, sogar wenn sie im Entscheidungstext nicht wörtlich auftreten. Gelegentlich tauchen außerdem alltagssprachliche Begriffe (bis hin zu Funktionswörtern wie *soweit* und *oder*) als Indexterme auf.[11] Die in

[11] Es sind weder mögliche Indexterme noch Regeln für deren Normalisierung vorgegeben. So finden sich neben Fällen, in denen komplexe Ausdrücke vollständig aus dem Urteilstext ins Regi-

Tabelle 2.12 angeführten Zahlen sind somit nur eingeschränkt aussagekräftig, und es ist davon auszugehen, dass bei einer detaillierteren Untersuchung (die wir allerdings im Rahmen dieser Arbeit nicht durchführen werden) auf einzelne Dokumente bezogen deutlich abweichende Übereinstimmungswerte festzustellen wären.

## 2.5 Fazit

Bei begriffsbestimmenden Aussagen in Gerichtsurteilen handelt es sich nur zu einem Teil um vollwertige Definitionen im Sinne der Äquivalenz- und Essentialitätsbedingung (vgl. 1.2.1). Auch das in der Terminologiewissenschaft als prototypisch für Definitionen betrachtete Formulierungsmuster *Ein A ist ein B, das C* wird bei weitem nicht in allen dieser Aussagen verwendet. Begriffsbestimmungen erfüllen unterschiedliche Aufgaben im Rahmen der Entscheidungsbegründung. Neben umfassenden Kerndefinitionen finden sich elaborierende Feststellungen, mit denen angepasst an den konkreten Fall und den argumentativen Zusammenhang partielle semantische Festlegungen und Präzisierungen getroffen werden oder auch erläuternde Zusatzinformation zu einem Terminus angegeben wird. In diesem erweiterten Sinne definitorische Information kann nicht aufgrund strikter Kriterien charakterisiert werden, sondern durch ein Bündel graduierbarer Merkmale. Ob eine Äußerung demnach als definitorisch anzusehen ist, kann zwar unter Umständen nur im Kontext und unter Einbeziehung von Welt- und Fachwissen beurteilt werden. Dennoch haben wir in einer Annotationsstudie zwischen einem juristisch ausgebildeten Annotator und einem Nicht-Juristen eine insgesamt befriedigende Übereinstimmung bei der Identifikation definitorischer Information im weiteren Sinne festgestellt. Eine genauere Analyse der Annotationsergebnisse zeigt außerdem, dass die Übereinstimmung zwischen den Annotatoren einen als substantiell zu bewertenden Grad erreicht, wenn besonders problematische Einzeldokumente außer Betracht gelassen werden. Es ist also anzunehmen, dass stilistische Eigenarten einzelner Urteilstexte die Identifikation von definitorischer Information erleichtern oder erschweren können.

Die zum Ausdruck von Definitionen in Urteilsbegründungen verwendeten Formulierungsmuster sind – insbesondere innerhalb der wichtigen Gruppe der prädikatbasierten Definitionen – vielfältiger als in bisherigen Untersuchungen zur Form von Definitionen festgestellt, jedoch ist eine Systematisierung der Ausdrucksmittel nach semantisch-funktionalen Gesichtspunkten möglich. Da-

ster übernommen wurden (z.B. *sich dem Antritt der erkannten Freiheitsstrafe entziehen*) auch Fälle, in denen jeder einzelne Termbestandteil in einer Grundform als eigenes Schlagwort verzeichnet wurde (z.B. *gleichartig* und *Grundsatz* für *Grundsatz der Gleichartigkeit*)

bei zeigt sich, dass Kopula und Klassifikationsverben sowie Prädikate, die explizit auf die Bedeutungsrelation Bezug nehmen, hauptsächlich in Kerndefinitionen verwendet werden. Dagegen treten Ausdrücke mit Bezug zum juristischen Prüfungs- und Argumentationsverlauf (eine in sich besonders stark ausdifferenzierte Klasse von Definitionsprädikaten) häufiger in elaborierenden Definitionen auf.

Im weiteren Verlauf der Arbeit werden wir untersuchen, wie die bisher identifizierten definitionstypischen Formulierungen als Suchmuster für die automatische Erkennung von Definitionen genutzt werden können, welche Arten linguistischer Information für eine zuverlässige Erkennung benötigt werden und wie erkannte Definitionen anhand ihrer sprachlichen Struktur analysiert und weiterverarbeitet werden können.

# Kapitel 3

# Textbasierter Informationszugriff, Definitionsextraktion, juristische Systeme

Aufgrund der stetig wachsende Menge global verfügbarer Information stoßen klassische Verfahren der Wissenserschließung und -verwaltung (etwa durch manuelle Katalogisierung und bibliothekarische Aufbereitung) seit langem an ihre Grenzen. Gleichzeitig ist die Entwicklung von Methoden zum automatischen Informationszugriff zu einem zentralen Thema der informationstechnologischen Forschung geworden.

Auch in den Rechtssystemen moderner Rechtsstaaten werden enorme Informationsbestände geschaffen, die so gut wie ausschließlich in textueller Form vorliegen. Zwar werden nur in einem relativ geringen Teil der bei Gerichten anhängigen Fälle Dokumente veröffentlicht. So werden zur Zeit bei deutschen Gerichten jährlich etwa 4 Millionen Fälle erledigt, die Zahl der veröffentlichten Entscheidungstexte liegt dagegen bei etwa 30 000 pro Jahr.[1] Allerdings kann auch diese Zahl mit traditionellen Methoden nur noch unter immensem Arbeitsaufwand dokumentiert und erschlossen werden. Wie in Kap. 1 näher erläutert bleiben zudem auch ältere Entscheidungstexte – nicht zuletzt wegen des in ihnen entwickelten begrifflichen Wissens – für die Beurteilung neuer Fälle potentiell relevant. So ergibt sich ein insgesamt erheblich größerer Dokumentenbestand, auf den im Rahmen der juristischen Praxis regelmäßig zugegriffen werden muss. Dieser ist inzwischen zu einem beträchtlichen Teil elektronisch verfügbar. Marktführer in der elektronischen Dokumentation deutscher Rechtsprechung ist die Firma *juris*, die auf ein bereits 1973 begonnenes Projekt des Bundesjustizministeriums zurückgeht. In ihren Beständen sind

[1] Quelle: Ralph Dornis, *juris GmbH*, persönliches Gespräch. In einer Untersuchung zur Publikation deutscher Gerichtsentscheidungen in den Jahren 1987-1993 schätzt Reinhard Walker die "Publikationsdichte" der Entscheidungen deutscher Gerichte in diesem Zeitraum mit 0,46% auf einen noch etwas geringeren Anteil, geht allerdings von einer deutlich steigenden Tendenz aus (Walker (1998), 76ff.). Die erhebliche Differenz zwischen erledigten Fällen und veröffentlichten Entscheidungen erklärt sich u.a. dadurch, dass ein großer Teil von Verfahren durch einen Vergleich oder durch Klagerücknahmen zum Abschluss kommt. Auch von den mit einem Urteil abgeschlossenen Verfahren wird allerdings bei weitem nicht immer der Entscheidungstext veröffentlicht.

über 800 000 Entscheidungstexte recherchierbar (mit einem Zuwachs von etwa 27 000 Entscheidungen z.B. im Jahr 2004). In den deutschsprachigen Entscheidungssammlungen der international agierenden Anbieter *LexisNexis* und *WestLaw* finden sich ca. 520 000 bzw. 215 000 Entscheidungen. Die Möglichkeiten zur automatischen tieferen Erschließung der Information in solchen Sammlungen sind jedoch bisher kaum systematisch untersucht worden.

Wir stellen im Folgenden den Stand der für das Thema der automatischen Identifikation und Analyse von Definitionen in Entscheidungstexten relevanten Forschungen dar. Wir geben zunächst einen knappen Überblick über Ansätze zum *textbasierten automatischen Informationszugriff* im allgemeinen (3.1) und gehen dann ausführlicher auf das im Rahmen dieser Arbeit besonders relevante Thema *Identifikation und Verarbeitung von Definitionen* ein (3.2). Im Anschluss wenden wir uns dem Stand der Forschung zu juristischen Anwendungen moderner Informationstechnologie zu (3.3). Abschließend diskutieren wir einige Besonderheiten der Rechtsdomäne, die die im Vergleich zu anderen Bereichen noch relativ geringe Bedeutung von Techniken des textbasierten Informationszugriffs in diesem Zusammenhang erklären können (3.4).

## 3.1 Automatischer textbasierter Informationszugriff

Die Nutzung von Computertechnologie zur automatischen Erschließung großer Dokumentensammlungen ist mindestens seit den 1960er Jahren als ein eigenes Forschungsgebiet der Informatik etabliert (Becker und Hayes (1963); Salton (1968); Kent (1971); Spärck Jones (1972)). Erst mit der Verfügbarkeit umfangreicher maschinenlesbarer Textbestände und leistungsfähiger Hardware konnten jedoch (im wesentlichen seit Anfang der 1990er Jahre) systematisch verschiedene Szenarien für den *textbasierten Informationszugriff*[2] erforscht und entsprechende Lösungen implementiert werden. Der Fokus der Forschung lag dabei auf drei Anwendungsfeldern, die wir im Folgenden näher betrachten werden: In 3.1.1 beschreiben wir Aufgabenstellung und Stand der Technik im *Information Retrieval*, dem automatischen Auffinden von Dokumenten zu Suchanfragen. In 3.1.2 diskutieren wir die – verwandte, aber komplexere – Aufgabe der *Information Extraction*, des gezielten Zugriffs auf bestimmte Arten von In-

[2] Auch wenn neu entstehendes Wissen wohl zu einem beträchtlichen (und vermutlich wachsenden) Anteil multimedial, also nicht allein sprachlich kodiert ist, liegt der Schwerpunkt der meisten Forschungen bisher auf der Nutzbarmachung textueller Informationsbestände. Für Anwendungen im juristischen Bereich stellt diese Fokussierung aufgrund des absoluten Vorrangs der Schriftlichkeit vor anderen Kommunikationswegen keine Einschränkung dar.

formation in Textbeständen. In 3.1.3 gehen wir auf Systeme zur automatischen Fragebeantwortung (*Question Answering*) ein.

Für alle drei Bereiche sind inzwischen generische Architekturen beschrieben, und es besteht Einigkeit über die Grundlagen der Evaluation implementierter Systeme. Hierfür wurden in jedem der Bereiche eine Anzahl von Referenzressourcen erstellt, anhand derer in regelmäßigen Wettbewerben die Performanz verschiedener Systeme ermittelt und verglichen wird.

Daneben werden aber auch verschiedene Aufgabenstellungen betrachtet, die nicht (oder zumindest nicht vollständig) einem der genannten Felder zuzuordnen sind und nicht im gleichen Maße paradigmatisch erforscht werden. Auf einige Beispiele gehen wir in 3.1.4 kurz ein.

### 3.1.1 Information Retrieval

Zielsetzung im Information Retrieval (IR) ist das *Auffinden* von Informationen in großen Dokumentenbeständen aufgrund von Suchanfragen. Typischerweise wird die Suchanfrage durch den Benutzer direkt in Form von Suchbegriffen spezifiziert. Im Regelfall versucht das System auf deren Basis ganze Dokumente zu identifizieren, je nach Anwendung und Aufbau der Textgrundlage kann aber auch nach kürzeren Textpassagen gesucht werden. Das Suchergebnis wird dem Benutzer in Form einer (unter Umständen nach Relevanz sortierten) Liste mit Verknüpfungen zu den aufgefundenen Texten präsentiert, die durch kurze, informative Textauszüge ergänzt sein können.

Die bekanntesten Beispiele für erfolgreich eingesetzte IR-Systeme stellen Internet-Suchmaschinen wie Google und Microsoft Bing dar. Auch die Suche nach technischer Dokumentation (z.B. in Online-Hilfesystemen bei Softwareprodukten), die Suche in digitalen Bibliotheken oder die Suche nach Patientenakten in Krankenhäusern bzw. Fallakten in Kanzleien sind wichtige Anwendungsbereiche des IR. Bei den Suchportalen elektronischer Entscheidungssammlungen wie *juris* handelt es sich ebenfalls um typische IR-Systeme. Wir gehen hier kurz auf die wichtigsten Aspekte einer generischen IR-Architektur ein. Dabei orientieren wir uns an Baeza-Yates und Ribeiro-Neto (1999).

#### (a) Generische IR-Architektur

Die Architektur eines generischen IR-Systems ist in Abb. 3.1 dargestellt. Kernkomponente eines IR-Systems ist der aus der Textbasis erzeugte *Index*. Er beinhaltet Repräsentationen der Dokumente (oder kleinerer adressierbarer Einheiten, z.B. Abschnitte oder Sätze) in der Textbasis. Im einfachsten Fall verzeichnet er für jedes Wort in der Textbasis alle Dokumente, in denen dieses auftritt. Ein solcher *inverted file index* ermöglicht ein direktes Auffinden aller Tref-

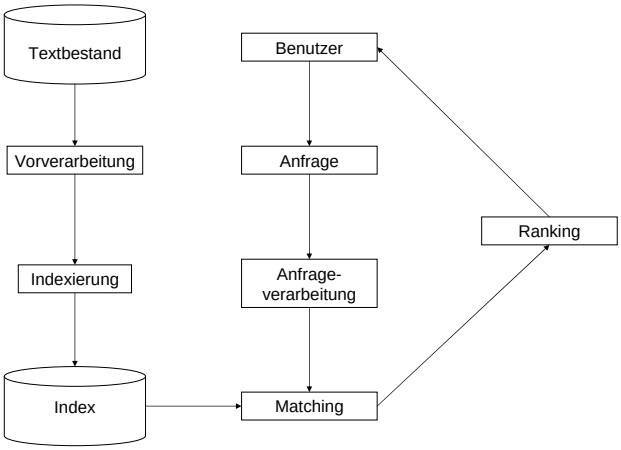



Abbildung 3.1: Generische IR-Architektur (vgl. Baeza-Yates und Ribeiro-Neto (1999), 10)

fer zu einem Suchbegriff. Um die Relevanz der erzielten Treffer zu erhöhen, können im Rahmen der Vorverarbeitung der Textbasis (etwa aufgrund der Dokumentstruktur oder unter Nutzung vorgegebener Schlagwortkataloge) einzelne Wörter als Indexbegriffe ausgewählt bzw. ausgeschlossen oder modifiziert (z.B. durch linguistische Verarbeitung normalisiert) werden. Weiterhin können anstatt eines *inverted file index* andere Indexstrukturen (z.B. Baumstrukturen) erzeugt werden, mit denen sich die Effizienz bei bestimmten Anfragen steigern läßt.

Vom Benutzer gestellte Suchanfragen müssen im Rahmen der Anfrageverarbeitung der Dokumentrepräsentation im Index angepasst werden, insbesondere müssen Suchbegriffe entsprechend der Vorverarbeitung der Textbasis normalisiert werden. Zusätzlich werden von manchen Systemen Veränderungen (z.B. Ergänzungen, sog. *query expansions*) an der Suchanfrage vorgenommen, um mehr oder hochwertigere Treffer zu erhalten.

In der Matching-Phase wird dann auf den Index zugegriffen, um für die Suchanfrage relevante Dokumente zu identifizieren. Man unterscheidet dabei verschiedene Ansätze (sog. *Retrieval Modelle*). In der Praxis kommen vor allem das *Boolsche Modell* und das *Vektorraum-Modell* zur Anwendung. Das Boolsche Modell erlaubt Verknüpfungen von Suchbegriffen mit logischen Operatoren (meistens *und*, *oder* sowie *nicht*). Es hat somit eine für den Benutzer direkt einsichtige Semantik und läßt sich zudem direkt durch mengentheoretische Operationen auf *inverted file indices* umsetzen.

Für das Vektorraummodell (Salton u. a. (1975)) werden sowohl Dokumente als auch Anfragen als Vektoren repräsentiert, die im einfachsten Fall für jedes Wort die Anzahl der Auftreten verzeichnen. Diese Repräsentationsform abstrahiert somit über die Reihenfolge der Wörter in einem Dokument und gewichtet häufige Wörter höher als seltene. Erweiterungen des Vektorraummodells nutzen Gewichtungen, die sich nicht allein auf Häufigkeiten im Einzeldokument stützen, sondern beispielsweise generell besonders häufige Wörter wie *ist* oder *der* geringer gewichten als Inhaltswörter mit vielen Auftreten in einem spezifischen Einzeldokument (wie etwa *Fundament* in der in 1.3.1 analysierten Textpassage).

Für jeden Dokumentvektor wird dann die Entfernung zum Anfragevektor ermittelt. Diese wird als Maß für die Ähnlichkeit zwischen der Anfrage und dem jeweiligen Dokument interpretiert. Entsprechend werden diejenigen Dokumente zurückgegeben, deren Vektoren die geringste Distanz zum Anfragevektor aufweisen. Das Vektorraum-Modell erlaubt verschiedenste Umsetzungen – die Auswahl der als Dimensionen dienenden Basiselemente, die Ermittlung der Werte für jede Dimension sowie der Vergleich zwischen Dokumentvektoren und Anfragevektor können auf jeweils unterschiedliche Weise realisiert werden. Die Auswahl der im Rahmen einer Anwendung optimalen Parametrisierung kann deshalb einen erheblichen Entwicklungsaufwand erfordern. Zudem ist die Semantik von Anfragen im Vektorraum-Modell für den Benutzer weit weniger transparent als bei der Suche mit Boolschen Operatoren. Diese Nachteile werden jedoch in vielen Fällen durch den Zugewinn an Flexibilität aufgewogen, der daraus resultiert, dass das Vektorraum-Modell anstelle einer binären Unterscheidung zwischen Treffern und irrelevanten Dokumenten ein gradiertes Relevanzmaß zur Verfügung stellt.

Auf dieses kann direkt zur Ordnung der Suchergebnisse für die Präsentation zugegriffen werden. Ein Ranking kann sich zudem – auch wenn ein Bool'sches Retrieval Modell zum Einsatz kommt – auf andere Eigenschaften der Suchergebnisse (z.B. Position im Ursprungsdokument, publizierende Instanz) oder externe Informationsquellen stützen. Bekanntestes Beispiel ist die Auswertung der Verlinkungs-Struktur des WWW durch Google, vgl. Brin und Page (1998).

In wissenschaftlichen Dokumentsammlungen – wie CiteSeer, vgl. Giles u. a. (1998) – werden zum Beispiel Zitationsgraphen exploriert.

#### (b) Nutzung linguistischer Information im IR

Die wichtigsten Phasen, in denen in der in Abb. 3.1 skizzierten Architektur linguistische Information genutzt werden kann, sind die Dokumentenvorverarbeitung und Indizierung sowie die Anfrageverarbeitung (vgl. auch Fliedner (2007), 11).

In aktuell praktisch eingesetzten Systemen werden oft vor der Indizierung der Textbasis (und entsprechend auch aus den vom Benutzer gestellten Anfragen) sogenannte Stoppwörter ausgefiltert, die aufgrund ihrer Verteilung (Vorkommen in allen oder fast allen Dokumenten) oder sprachlichen Funktion (z.B. Funktionswörter) nicht zur inhaltlichen Unterscheidung zwischen Dokumenten beitragen. Die verbleibenden Indexterme werden zudem meist durch heuristische Regeln zur Entfernung von Flektionsmerkmalen (sog. *Stemming*) oder durch vollständige morphologische Analyse (unter Umständen einschließlich einer Kompositazerlegung) normalisiert. Auch die bereits angesprochenen *query expansion*-Techniken sind typischerweise linguistisch motiviert. Unter Nutzung allgemeinsprachlicher Ressourcen (z.B. des maschinenlesbaren englischen Thesaurus Wordnet, Fellbaum (1998)) oder fachsprachlicher Thesauri werden Suchbegriffe um Synonyme ergänzt oder durch Oberbegriffe ersetzt, um größere Treffermengen zu erhalten. In Forschungsprototypen wird darüber hinaus mit Indizes auf der Grundlage weitergehender linguistischer Analysen (etwa zu Kollokationen, Phrasen oder umfassenderen syntaktischen Strukturen), mit interaktiven Methoden der *query expansion* in Dialogform sowie mit mehrsprachiger *query expansion* zur Suche in multilingualen Dokumentsammlungen experimentiert.

Bisherige Forschungen haben kein einheitliches Bild zum tatsächlichen Nutzen linguistisch motivierter Erweiterungen in IR Systemen ergeben. Es erscheint plausibel, anzunehmen, dass dieser in hohem Maße von den Eigenschaften des jeweiligen Textbestandes (insbesondere von dessen Sprache) abhängt.

#### (c) Evaluationsstandards

Als Grundlage der Evaluation von IR-Systemen haben sich die aus der Signalverarbeitung übernommenen Standardmaße *Präzision* und *Recall* etabliert (Cleverdon u. a. (1966); Cleverdon (1970)):

**Präzision** erfasst den Anteil relevanter Treffer (d.h. *true positives*) am Suchergebnis und berechnet sich folgendermaßen:

$$p = \frac{|\text{Relevante Treffer}|}{|\text{Alle Treffer}|}$$

**Recall** bezeichnet den Anteil der relevanten Treffer an der Gesamtzahl der relevanten Dokumente in der Textgrundlage (und erfasst somit indirekt die *false negative*-Rate):

$$r = \frac{|\text{Relevante Treffer}|}{|\text{Alle Dokumente}|}$$

Als Kombination beider Größen wird der als *f-score* bezeichnete harmonische Mittelwert angegeben:

$$f = \frac{2pr}{p+r}$$

Zwischen Präzision und Recall besteht in vielen Fällen ein inverser Zusammenhang (Cleverdon (1972)): Verbesserungen der Präzision führen zu einem Verlust an Recall, während ein höherer Recall oft nur durch einen Verzicht auf Präzision zu erreichen ist. Verschiedene Anwendungsszenarios können dabei eine unterschiedliche Abwägung zwischen beiden Größen erfordern. So kann bei der Suche in großen redundanten Dokumentbeständen eine Optimierung der Präzision zu Ungunsten des Recall sinnvoll sein, die bei einer kleineren Textgrundlage ohne Redundanzen das Suchverfahren unbrauchbar machen würde. Um dieser Abwägung Rechnung zu tragen, kann bei der Berechnung des *f-score* ein Gewichtungsfaktor $\beta$ berücksichtigt werden:

$$f_\beta = (1+\beta^2)\frac{pr}{\beta^2 p + r}$$

Für die Ermittlung der angesprochenen Performanzmaße wird eine aufgearbeitete Textbasis und eine Menge von Musteranfragen benötigt. Zu jedem der in der Textbasis enthaltenen Dokumente (bzw. sonstigen Zieleinheiten der Informationssuche) muss bekannt sein, für welche der Musteranfragen es jeweils einen relevanten Treffer darstellen würde. Die Erstellung solcher Referenzressourcen erfordert die Sichtung einer Vielzahl von Dokumenten und ist daher mit großem Arbeitsaufwand verbunden. Es ist zudem unklar, wie weit Evaluationsergebnisse verschiedener IR-Systeme auf unterschiedlichen Referenzressourcen sinnvolle Vergleiche zwischen den Systemen zulassen.

In den 1990er Jahren wurden daher zur Evaluation von IR-Systemen sogenannte *shared task*-Konferenzen und -Wettbewerbe eingerichtet. Bei diesen

wird zentral ein detailliert aufgearbeitetes und einheitlich formatiertes Textkorpus mit Musteranfragen und Relevanzangaben auf der Basis kontrollierter Mehrfachannotationen zur Verfügung gestellt. Teilnehmende Forschungsgruppen erhalten zunächst Trainingsdaten als Grundlage für die Systementwicklung und ermitteln die Performanz ihrer Systeme dann anhand von Testdaten, die erst nach Abschluss der Entwicklungsfrist verfügbar gemachten werden. Die wichtigsten solchen *shared task*-Reihen sind die 1992 vom US-amerikanischen *National Institute for Science and Technology (NIST)* initiierte *Text Retrieval Conference (TREC)*, auf europäischer Seite das *Cross Linguistic Evaluation Forum (CLEF)* mit einem besonderen Schwerpunkt auf multilingualem Informationszugriff und in Japan die Workshops auf der Basis der *NTCIR (National Institute of Informatics Test Collection for IR Systems)*. Ein wichtiges fachspezifisches Referenzkorpus (ohne zugehörige Konferenzreihe) stellt die *Cystic Fibrosis Collection* dar, in der 1239 Dokumente aus der MEDLINE-Datenbank der *US National Library of Medicine* mit gestuften Relevanzwerten für 100 von medizinischen Experten formulierte Suchanfragen annotiert sind.

Die Nutzung der genannten Ressourcen und die verschiedenen *shared task*-Wettbewerbe haben zu großen Fortschritten in der Entwicklung von IR-Systemen (und, wie in den folgenden Abschnitten noch darzulegen ist, auch von Systemen zum Informationszugriff im allgemeinen) geführt. Zum einen haben sie dazu beigetragen, rigorose quantitative Evaluationsmethoden zu etablieren. Sie haben damit die Grundlage für transparente Vergleiche zwischen verschiedenen Ansätzen und Implementierungen geschaffen. Zum anderen erhalten Forschungsprojekte durch die Verfügbarkeit aufgearbeiteter, einheitlich formatierter und systematisch annotierter Entwicklungs- und Testdaten die Möglichkeit, einen größeren Teil ihrer Ressourcen auf die Konzeption und Entwicklung der eigentlichen Kernkomponenten ihrer Systeme zu verwenden. Allerdings hat die Fokussierung auf vorgegebene Daten auch dazu geführt, dass spezifische Anforderungen einzelner Domänen und Textsorten in der Forschung bisher größtenteils wenig Beachtung erfahren haben, wenn sie (wie etwa Gerichtsentscheidungen) von den *shared tasks* nicht abgedeckt waren.

### 3.1.2 Information Extraction

Während durch IR Systeme aufgrund von Suchanfragen ganze Dokumente (oder zumindest größere Textpassagen) in einer Textbasis aufgefunden werden sollen, greifen *Information Extraction (IE)*-Systeme direkt auf einzelne, nach thematischen Kriterien ausgewählte Informationsbestandteile zu. Zielsetzung ist typischerweise die Überführung textueller Information in ein strukturiertes Format, etwa die Übernahme in Berichtsschablonen oder auch Tabellen einer relationalen Datenbank.

**(a) Generische IE-Architektur**

Abb. 3.2 gibt die Architektur eines generischen IE-Systems wieder.

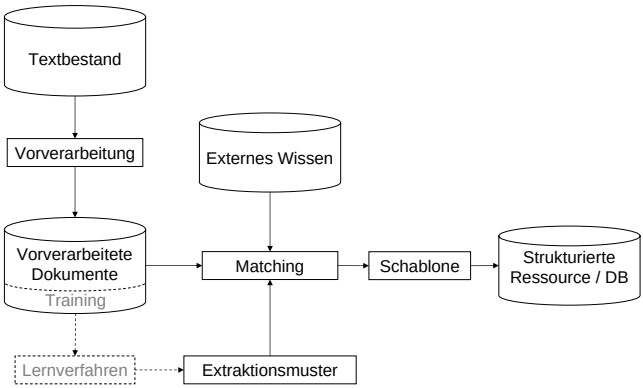



Abbildung 3.2: Generische IE-Architektur (vgl. Moens (2006), 37; Turmo u. a. (2006), 8)

Die Textgrundlage und der zu befriedigende Informationsbedarf (der sich in den Ziel-Ergebnisstrukturen widerspiegelt) definieren die von einem IE-System zu erfüllende Aufgabe. Zentraler Systembestandteil ist die Extraktionsgrammatik. Diese enthält Extraktionsregeln oder -muster, die entweder typische sprachliche Realisierungsmöglichkeiten für die gesuchte Information selber oder typische sprachliche Umgebungen des gesuchten Informationstyps beschreiben. Zudem legen die Extraktionsregeln fest, wie die identifizierten Textbestandteile auf die Ausgabestrukturen abzubilden sind.

Die Struktur der Extraktionsregeln hängt dabei zum einen von der Art der Zielstrukturen, zum anderen von der Art der eingesetzten (insbesondere sprachtechnologischen) Vorverarbeitung und der im Extraktionsprozess verfügbaren zusätzlichen Ressourcen ab. In groben Zügen kann die Identifikation der Zielinformation – und dementsprechend auch die Extraktionsgrammatik – in vielen Anwendungsszenarios in hierarchisch aufeinander aufbauende Einzelaufgaben gegliedert werden. Die in der Literatur diskutierten Einteilungen differie-

ren zwar im Einzelnen, unterscheiden aber meist in etwa folgende Teilbereiche (vgl. z.B. Moens (2006); Xu (2007), 50):

- **Named Entity-Erkennung:** Die Erkennung und Klassifikation von Eigennamen und ähnlichen Ausdrücken (etwa Firmenbezeichnungen, Bezeichnungen von Geldbeträgen oder – in der biomedizinischen Domäne – Bezeichnungen für Proteine). Die Struktur solcher Ausdrücke folgt oft Eigengesetzlichkeiten außerhalb der allgemeinsprachlichen Grammatik, interagiert aber andererseits mit der linguistischen Struktur des umgebenden Textes. Ihre Erkennung wird deshalb in manchen Systemen in der Vorverarbeitung / linguistischen Analyse integriert.

- **Koreferenzauflösung:** Die Identifikation von Ketten von Ausdrücken, die jeweils dasselbe Objekt bezeichnen. Dies umfasst nicht nur die Auflösung von Anaphern (also z.B. Pronomen, mit denen auf vorerwähnte Entitäten Bezug genommen wird), sondern beispielsweise auch die Erkennung unterschiedlicher Versionen desselben Namens oder von Namen und Beschreibungen mit identischer Referenz.

- **Relationserkennung:** Die Erkennung und Einordnung von einfachen, generischen oder domänentypischen Relationen zwischen Objekten (die in der Regel zuvor als Named Entities identifiziert wurden).

- **Szenario- / Ereigniserkennung:** Die Erkennung komplexerer Muster, die mehrere Relationen sowie Relationen zwischen diesen und weiteren Entitäten umfassen und bspw. Abläufe oder Kausalitäten beschreiben.

Die Extraktionsregeln können von Domänenexperten handkodiert oder mit Techniken des maschinellen Lernens automatisch erworben werden. Der expertenbasierte Ansatz erlaubt in der Regel sehr präzise Extraktionsergebnisse (vgl. Appelt und Israel (1999), 10). Jedoch ist die manuelle Erstellung von Extraktionsgrammatiken sehr aufwändig. Insbesondere in Fällen, in denen der gesuchte Informationstyp eine Vielzahl nicht-standardisierter oder unspezifischer Ausdrucksmöglichkeiten zulässt, ist zudem nicht davon auszugehen, dass diese von Experten vollständig identifiziert werden können. Als Alternative zur manuellen Spezifikation werden daher maschinelle Lernverfahren zum automatischen Erwerb von Suchmustern angewandt. Der direkte automatische Erwerb von Extraktionsregeln (z.B. Soderland (1999)) setzt die Verfügbarkeit annotierter Trainingsdaten voraus. Die Präzision der mit den erworbenen Regeln erzielten Extraktionsergebnisse ist außerdem tendenziell niedriger als bei manuell kodierten Extraktionsgrammatiken. Jedoch können große Zahlen von Extraktionsregeln generiert und potentiell auch solche Muster erfasst werden, die von

Domänenexperten nicht als charakteristische Realisierungen des gesuchten Informationstyps angesehen würden. Neben Methoden zum direkten automatischen Erwerb von Extraktionsregeln kommen (vor allem in jüngeren Ansätzen) vermehrt *Bootstrapping*-Verfahren zum Einsatz (z.B. Riloff und Jones (1999); Xu (2007)). Dabei werden ausgehend von einer geringen Anzahl von Schlüsselwörtern aus bekannten Instanzen des gesuchten Informationstyps (sog. *Seeds*) Formulierungsmuster in einem großen Textbestand identifiziert, die wiederum zur Gewinnung neuer *Seeds* genutzt werden. Dieser Zyklus wird wiederholt, bis eine definierte Abbruchbedingung erfüllt ist.

*Bootstrapping*-Verfahren vermeiden die Abhängigkeit von der Verfügbarkeit und Qualität manuell annotierter Trainingsdaten. Durch die Wahl der Abbruchbedingung sowie die Nutzung unterschiedlicher Verfahren zur Auswahl der neu gewonnenen Formulierungsmuster und *Seed*-Ausdrücke erlauben sie zudem eine flexible empirische Anpassung an unterschiedlich beschaffene Textgrundlagen. Sie erleichtern somit gegenüber manuellem Regel-Engineering und gegenüber anderen automatischen Lerntechniken die Übertragung von IE-Systemen zwischen verschiedenen Domänen.

#### (b) Einsatz linguistischer Technologie

Im Gegensatz zur Zielsetzung reiner IR-Systeme erfordern typische IE-Aufgaben eine sehr viel weitreichendere Berücksichtigung der strukturellen Zusammenhänge zwischen den elementaren Informationseinheiten in der Textgrundlage. Zudem erlauben insbesondere komplexere Zusammenhänge, auf die etwa die Szenario- bzw. Ereigniserkennung abzielt, meist eine Vielzahl im Detail unterschiedlicher, in semantischer Hinsicht jedoch weitgehend gleichwertiger sprachlicher Realisierungsmuster. Verglichen mit typischen IR-Systemen machen IE-Systeme daher in relativ hohem Maße Gebrauch von linguistischen Analysetechniken.

Über die (auch in vielen IR-Systemen genutzten) Schritte der Dokumentstruktur-Analyse, Tokenisierung und morphologischen Analyse (bzw. des Stemming) hinaus werden in den meisten IE-Systemen zumindest partielle syntaktische Strukturen ermittelt. Häufig werden Nominalgruppen zusammengefasst (*NP-Chunking*), einige Systeme erzeugen vollständige syntaktische Parses. In neueren Ansätzen haben sich die Verwendung von Dependenzanalysen (vgl. 4.2.3) und die Kombination verschiedener, jeweils eingeschränkter aber zuverlässiger Analysetechniken als besonders erfolgreich erwiesen.

Art und Umfang der in einem IE-System genutzten linguistischen Vorverarbeitung bestimmen darüber, welche Strukturen in den Zugriffsphase für den Abgleich mit den Extraktionsregeln verfügbar sind. Sie haben somit direkten Ein-

| | *Named Entities* | *Koreferenz* | *Template-Elem.* | *Relationen* | *Szenarios* |
|---|---|---|---|---|---|
| *Präzision* | 0,95 | 0,69 | 0,87 | 0,86 | 0,65 |
| *Recall* | 0,92 | 0,56 | 0,87 | 0,67 | 0,42 |

Tabelle 3.1: Präzision und Recall der besten MUC-7-Systeme, vgl. Chinchor (1998)

fluss auf das Format dieser Regeln. Beschränkt sich die linguistische Verarbeitung auf lexikalisch-morphologische Normalisierung oder Phrasen-Chunking, werden Suchmuster meist als lineare Wort-, Wortart- oder Phrasensequenzen formuliert. Bei Extraktionsregeln für tiefer verarbeitete Daten (also zum Beispiel Dependenz- oder Prädikat-Argument-Strukturen) werden in der Literatur verschiedene Modelle – von einfachen Tupeln aus Prädikat und direkten Dependenzen bis zu beliebigen Unterbäumen – unterschieden.

#### (c) Evaluation

Auch für die Evaluation von IE-Systemen haben sich für Wettbewerbsreihen entwickelte Ressourcen und Methoden als Quasi-Standard etabliert. Im Rahmen der von 1987 bis 1998 veranstalteten *Message Understanding Conferences (MUCs)* wurden verschiedene Anwendungsdomänen erprobt. Die Testdaten entstammen insgesamt fünf Domänen – *Schiffsmeldungen*, *terroristische Aktivitäten*, *Satellitenstarts*, *Joint Ventures / Mikroelektronik* und *Personalwechsel in Führungspositionen*. Für Trainingsdaten wurden zum Teil noch weitere andere Felder ausgewählt. Als Textgrundlage dienten Zeitungsartikel oder ähnliche kurze, faktenorientierte Meldungen (bei der abschließenden MUC-7 jeweils 100 Artikel in zwei Trainingskorpora und 100 in einem Testkorpus). Bewertet wurde die Performanz der teilnehmenden Systeme bezüglich der oben geschilderten Teilaufgaben und der weiteren Aufgabenstellung der *Template Element Extraction*. Bei dieser waren für bestimmte Klassen von Entitäten zuvor festgelegte Attribute aus dem Text zu extrahieren, also etwa für *Personen* neben dem Namen auch die *Nationalität* und ein eventueller *Titel*. Als Evaluationsmaße wurden den Aufgabenstellungen entsprechend angepasste Varianten der Größen Präzision und Recall verwendet. Zudem wurden auch Erweiterungen zur Einbeziehung partieller Treffer, zur Erfassung von *false positives* (sog. *overgeneration*) und der Gesamtfehlerrate (*fall out*) erprobt. Tabelle 3.1 enthält die

im Rahmen von MUC-7 von den jeweils besten Systemen erzielten Präzisions- und Recall-Werte.

Sehr gute Werte werden lediglich für die *Named Entity*-Erkennung erreicht. Für alle nicht allein auf einzelne Entitäten bezogenen Aufgaben fallen die Werte deutlich ab, wobei die Szenarioerkennung durch Tabelle 3.1 klar als schwierigste Aufgabe ausgewiesen ist.

Seit 1999 hat die *Automatic Content Extraction (ACE)*-Initiative die MUCs als wichtigstes Forum zur Evaluation und zum Performanzvergleich von IE-Systemen abgelöst. Die ACE-Definition der relevanten Teilaufgaben deckt sich nur teilweise mit der für die MUCs angenommenen (so umfasst die Aufgabenstellung *Entity Detection and Tracking (EDT)* Aspekte der *Named Entity*-Extraktion *und* der Koreferenzauflösung, und die *Temporal Expression Detection* (TERN) wurde im Rahmen der MUCs nicht untersucht). Zudem verwenden die ACE-Evaluationen ein komplexes Evaluationsmaß, dass durch die Akkumulation von Bonus- und Maluswerten über die gesamten Ergebnisse des Testlaufs eines Systems berechnet wird. Aus diesen Gründen ist ein direkter Vergleich mit den Evaluationsergebnissen der MUCs nicht möglich. Tendenziell bestätigen die bei den ACE-Evaluationen erhobenen Daten jedoch die auch aus Tabelle 3.1 ersichtliche Staffelung des Schwierigkeitsgrads der einzelnen IE-Teilaufgaben.

### 3.1.3 Question Answering

Ein drittes wichtiges Forschungsfeld im Bereich des automatischen Informationszugriffs ist das automatische Beantworten von Fragen durch sogenannte *Question Answering (QA)*-Systeme. Aufgabe eines QA-Systems ist es, auf natürlich-sprachliche Anfragen des Benutzers passende Informationen aus einer Textbasis als Antwort anzubieten. Im Idealfall werden präzise, fokussierte und nicht redundante Antworten in natürlicher Sprache gegeben. Bisherige Systeme beschränken sich größtenteils auf sogenannte *faktoide Fragen*, deren Gegenstand eng eingegrenzte Tatsachen sind und die sich daher mit einem einzigen Satz oder sogar einer einzigen Phrase beantworten lassen. In jüngerer Zeit haben aber auch sogenannte *long answer questions* ein höheres Maß an Aufmerksamkeit erfahren. Hierzu gehören u.a. die durch allgemeingültige deskriptive Aussagen über ein Konzept zu beantwortenden *Definitionsfragen*, auf die wir in 3.2.3 gesondert eingehen.

### (a) Generisches QA-System

Die in Abb. 3.3 dargestellte generische QA-Architektur spiegelt die im Vergleich zu IR und IE deutlich größere Komplexität dieser Aufgabenstellung wider.

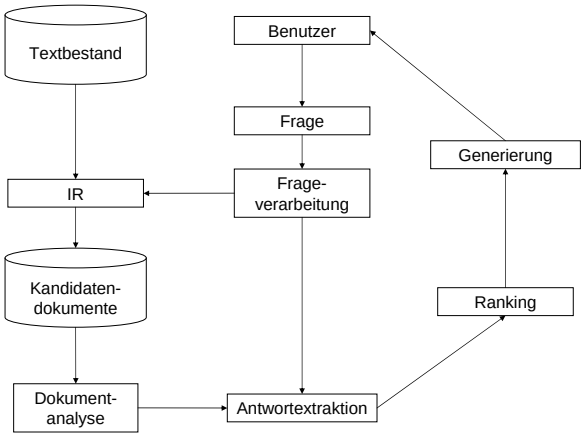



Abbildung 3.3: Generische QA-Architektur, vgl. Fliedner (2007), 22 und Hirschman und Gaizauskas (2001), 286

Neben der Verarbeitung der Textbasis und dem eigentlichen Informationszugriff stellen in QA-Systemen die Anfrageverarbeitung und die Antwortgenerierung eigene komplexe Teilaufgaben dar. Im Rahmen der Frageverarbeitung muss aus der vom Benutzer gestellten Frage auf die Art der angefragten Information (z.B. *ja/nein*-Entscheidung, Datum, Ort, Einzelfakt oder komplexer Zusammenhang) geschlossen werden. Außerdem muss der Frage Information entnommen werden, die geeignet ist, um antwortrelevante Passagen in der Textbasis zu identifizieren. In der Antwortgenerierungsphase muss aus den anhand der Benutzerfrage identifizierten Teilen der Textbasis eine syntaktisch wohlgeformte Äußerung erzeugt werden, die eine unter pragmatischen Gesichtspunkten zur Benutzerfrage passende Antwort darstellt (also insbesondere dem erfragten Antworttyp entspricht) und außerdem in inhaltlicher Hinsicht adäquat ist. Viele bisher umgesetzte QA-Systeme verzichten auf diesen letzten Verar-

beitungsschritt und präsentieren dem Nutzer direkt sehr eng eingegrenzte extrahierte Passagen aus der Textbasis.

Auch der eigentliche Informationszugriff ist in den meisten QA-Systemen aufwändiger als im IR oder der IE. In der Regel werden mit IR-Techniken anhand von Schlüsselbegriffen aus der Frage potentiell antwortrelevante Dokumente oder Passagen in der Textbasis aufgesucht. Die Treffermenge wird dann in einer Kaskade von Verarbeitungs- und Filterungsschritten weiter eingeengt.

Einige Systeme verwenden zusätzliche sog. *Antwortprojektions*-Verfahren. Dabei werden in externen Quellen (etwa in Datenbanken oder durch eine allgemeine Websuche) Antwortkandidaten gesucht und dann ähnliche Passagen in der eigentlichen Textbasis des Systems identifiziert (z.B. in Brill u. a. (2001) und Katz u. a. (2003)).

#### (b) Linguistische Verarbeitung

Linguistische Technologie kommt in QA-Systemen in allen Verarbeitungsschritten zum Einsatz (vgl. für eine detaillierte Analyse der Einsatzmöglichkeiten linguistischer Komponenten Fliedner (2007)). In der Frageverarbeitung wird auf syntaktische Analysen und semantische Klasseninformation (zumeist aus dem bereits erwähnten Thesaurus *Wordnet*) zurückgegriffen, um den Fragetyp und die mit der Frage verbundenen Einschränkungen möglicher Antworten zu ermitteln. In Frage-Antwort-Dialogen kann zudem der Fragekontext zu berücksichtigen sein, etwa wenn eine Frage anaphorische Elemente enthält. Bei der Identifikation von Dokumenten mit antwortrelevanter Information kommen häufig die in 3.1.1 angesprochenen linguistisch motivierten *query expansion*-Techniken zum Einsatz. Als Grundlage für die zielgerichtete Antwortextraktion und die sich eventuell anschließende Generierungsphase werden dann auf den identifizierten Dokumenten, einer gegenüber der gesamten Textbasis meistens um mehrere Größenordnungen kleineren Datenmenge, tiefere linguistische Analysen vorgenommen (Fliedner (2007), 247). Einige der erfolgreichsten aktuellen QA-Systeme arbeiten mit vollständigen syntaktischen Analysen (z.B. Cui u. a. (2004b)). Experimentiert wird weiterhin mit der Nutzung von Dependenzinformation (z.B. Katz und Lin (2003) und Fliedner (2007)), Prädikat-Argument-Strukturen (Harabagiu u. a. (2005); Moldovan u. a. (2002)) oder sogar logikorientierten semantischen Repräsentationen (etwa in Burhans (2002) und Friedland u. a. (2004)). In dieser Phase kommen zudem oft IE-Subsysteme zur Identifikation vordefinierter Informationstypen für bestimmte typische Fragen zum Einsatz. Soweit eine echte Antwortgenerierung stattfindet, stützt diese sich in der Regel auf die im Rahmen der Antwortextraktion ermittelten linguistischen Strukturen und wird mit Verfahren umgesetzt, die auch in Textgenerierungssystemen in der Realisierungsphase zum Einsatz kommen.

#### (c) Evaluation

Die bedeutendsten Foren für die Evaluation von QA-Systemen bilden die bereits in 3.1.1 genannten *shared task*-Wettbewerbsreihen TREC, CLEF und NTCIR, bei denen Ende der 1990-Jahre eigene QA-Tracks eingeführt wurden. Der Fokus dieser Wettbewerbe liegt auf der Beantwortung faktoider Fragen anhand von Zeitungsartikeln. Das wichtigste eingesetzte Performanzmaß ist die Akkuratheit eines QA-Systems, gemessen als der Anteil der korrekt beantworteten Fragen an allen gestellten Fragen. Die quantitative Bewertung der Performanz von QA-Systemen wirft jedoch insgesamt weit größere Probleme auf als im Falle von IR- oder IE-Systemen. So existieren in den meisten Fällen mehrere Formulierungsmöglichkeiten für die korrekte Beantwortung einer Frage sowie verschiedene Arten suboptimaler Antworten. Im Rahmen der genannten Wettbewerbe werden die Systemantworten daher manuell relativ feinkörnig klassifiziert. TREC lässt beispielsweise nicht nur *right* oder *wrong*, sondern auch *inexact* oder *unsupported* als Bewertungen zu, wenn Antworten neben korrektem noch irrelevantes Material enthalten bzw. korrekt, aber nicht auf die korrekte Textstelle gestützt sind. Damit ist jedoch nur ein geringer Teil der Dimensionen abgedeckt, in denen teilweise nur gradiert erfassbare Abweichungen von der / den optimalen Antworten möglich sind (zum Beispiel noch im Hinblick auf *Verständlichkeit* und *erkennbare Begründetheit*). Die Evaluationsansätze der bisherigen *shared task*-Wettbewerbe werden daher immer wieder kritisiert (Hirschman und Gaizauskas (2001), De Boni (2004)). Jedoch sind bisher auch keine Alternativen für eine quantitative Evaluation von QA-Systemen ausformuliert worden.

### 3.1.4 Weitere Forschungsthemen

Die drei bisher diskutierten Forschungsthemen IR, IE und QA sind gekennzeichnet durch relativ klar definierte Aufgabenstellungen und Anwendungsszenarios, weitgehende Einigkeit über die anzuwendende Architektur sowie etablierte quantitative Evaluationsstandards mit allgemein verfügbaren Referenzressourcen. Sie erschöpfen den Bereich *automatischer Informationszugriff* jedoch keineswegs. Weitere aktuelle Forschungsthemen sind beispielsweise:

- **Automatisches Zusammenfassen:** Die automatische Erstellung einer (in der Regel extraktiven) Zusammenfassung eines vorgegebenen Textes. Hauptforschungsfelder sind hierbei die Identifikation zu extrahierenden Materials im Originaldokument (v.a. anhand lexikalischer Merkmale und der Diskursstruktur) sowie die Erzeugung eines kohärenten Ausgabetextes (zum Stand der Forschung siehe z.B. Mani (2001) und Jones (2007)).

- **Ontologieextraktion:** Die automatische Erkennung und Extraktion von domänenspezifischen Konzepten und deren wichtigsten Attributen und Relationen (sog. *Domänenontologie*) auf der Grundlage eines größeren Bestandes an Fachtexten. Meist werden hierbei – im Anschluss an Hearst (1992) – relativ einfache Suchmuster zur Identifikation von Information des gesuchten Typs verwendet und mit statistischen Verfahren kombiniert, die auch bereits bekannte ontologische Information verwerten (vgl. etwa Lin und Pantel (2002); Cimiano (2006). Einen Überblick über den Forschungsstand bieten Buitelaar u. a. (2005)).

- **Text Mining:** Die automatische Entdeckung unbekannter Information in großen Textbeständen. Hierbei werden Aufgabenstellungen untersucht, die über die (oft ebenfalls unter dem Begriff Text Mining subsumierten) Teilaufgaben der IE hinausgehen. Es wird vor allem die Anwendbarkeit allgemeiner statistischer Verfahren zur Datenanalyse und -exploration auf textuelle Daten erprobt (vgl. etwa Kao und Poteet (2006) und Feldman und Sanger (2006)).

Diese Themen (und viele weitere im Bereich des textbasierten Informationszugriffs) weisen zwar einige Ähnlichkeiten und Überschneidungen mit den bisher besprochenen auf – etwa müssen für die Ontologiextraktion wie im Falle der Information Extraction Ausdrucksformen bestimmter Relationen identifiziert werden, und die Erstellung von Zusammenfassungen erfordert wie IR und QA die Identifikation relevanter Textpassagen. Andererseits fehlen beispielsweise bei der Entwicklung von Systemen zum automatischen Textzusammenfassen bisher verlässliche und allgemein anerkannte quantitative Kriterien für die Bewertung der Ergebnisse. In der Terminologie- und Ontologieextraktion (ebenso in vielen Bereichen des Text Mining) ist zudem vielfach kein direkter Bezug zu einem konkreten Anwendungsszenario gegeben. Untersucht wird vielmehr auf eher prinzipieller Ebene die Erzeugung oder Ergänzung von Wissensbeständen, die in erst noch zu bestimmenden Anwendungskontexten genutzt werden sollen. Aus diesen Gründen existieren in den genannten Bereichen keine allgemein akzeptierten generischen Architekturen. Verschiedene Systeme gehen oft von relativ unterschiedlichen Rahmenbedingungen aus und sind nicht direkt vergleichbar.

## 3.2 Definitionsextraktion

Auch die automatische Extraktion von Definitionen in Texten überschneidet sich in vieler Hinsicht mit anderen Forschungsthemen des textbasierten Informationszugriffs, ist jedoch keinem dieser Felder vollständig zuzuordnen. Ansät-

ze zur Extraktion und Verarbeitung von Definitionen in größeren Textbeständen wurden und werden in verschiedenen Kontexten verfolgt. Dabei sind auch die jeweils betrachteten Aufgabenstellungen und Zielsetzungen im einzelnen unterschiedlich. Wir gehen im Folgenden auf drei wichtige Aufgabenstellungen näher ein:

1. Eine relativ eng eingegrenzte Aufgabe ist die bloße *Identifikation definitorischer Textpassagen* in einer Dokumentensammlung ohne weitere Analyseschritte. Diese Zielsetzung verfolgen verschiedene Ansätze in der Terminologiewissenschaft und Computerlexikographie. Anwendungshintergrund ist hier die Erzeugung von Konkordanzen für den Terminologieexperten oder Lexikographen mit für die Begriffsbestimmung besonders relevanten Korpusbeispielen (3.2.1).

2. Eine zweite Aufgabenstellung, die in der Forschung verfolgt wird, ist die automatische *Erzeugung strukturierter Wissensressourcen* auf der Basis von Definitionen in Texten (3.2.2). Solche Ressourcen können dann beispielsweise als Domänenontologien zur Inferenzunterstützung genutzt werden. Diese Aufgabenstellung erhält dadurch eine besondere Komplexität, dass sie neben der Definitionsextraktion auch eine – unter Umständen sogar formale – Analyse der extrahierten Definitionen erfordert.

3. Bei der Beantwortung sogenannter *Definitionsfragen* durch QA-Systeme handelt es sich ebenfalls um eine komplexere Aufgabe. Solche Fragen des Typs *Wer / Was ist . . . ?* waren Bestandteil der QA-Tracks mehrerer TREC-Wettbewerbe. Ihre Beantwortung erfordert nicht nur das Auffinden einer gewissen Anzahl definitorischer Textpassagen, sondern auch deren Zusammenführung und Linearisierung in einer möglichst relevanten und nicht-redundanten Systemantwort (3.2.3).

Die erste und zweite Aufgabe weisen zwar offenkundig größere Überschneidungen mit dem Thema dieser Arbeit auf als die dritte. Die Beantwortung von Definitionsfragen ist jedoch – aufgrund der Einbindung in QA-Wettbewerbe – erheblich intensiver erforscht worden als die anderen beiden Aufgabenstellungen. Da wir zudem auf viele der im Definitions-QA umgesetzten Ansätze auch zur Ergebnisverbesserung in unserem eigenen Definitionsextraktionsverfahren zurückgreifen werden, gehen wir auch auf diese Thematik im Folgenden näher ein.

### 3.2.1 Identifikation definitorischer Textpassagen

Für die Lexikographie und Terminologiewissenschaft ist die Sammlung von Verwendungsbeispielen für Begriffe in Texten von zentraler Bedeutung. Mit der

Verfügbarkeit großer maschinenlesbarer Textbestände ist die Entwicklung automatischer Verfahren zur Suche nach aussagekräftigen Belegstellen in den Blickpunkt der lexikographischen und terminologiewissenschaftlichen Forschung gerückt. Aufgrund ihres explizit metasprachlichen Charakters kommt definitorischen Textpassagen dabei besonderes Interesse zu.

#### (a) Ansätze

Implementierte Ansätze zur textbasierten Identifikation von Definitionen für technisch-wissenschaftliche Fachtermini sind in Pearson (1998), Meyer (2001), Sierra und Alarcón (2002) und Barrière (2004b) beschrieben. Während die Untersuchungen dieser Autoren eine lexikographische Zielsetzung verfolgen, fanden in jüngerer Zeit auch in einigen Projekten zur hypermedialen Aufbereitung informativer Texte Forschungen zur Definitionsidentifikation für Fachtermini statt (mit dem Ziel der automatischen Erstellung verlinkter Glossare, Storrer und Wellinghoff (2006); Przepiorkowski u. a. (2007); Del Gaudio und Branco (2007); Westerhout und Monachesi (2007)).

In Tabelle 3.2 sind die genannten Ansätze zusammengefasst (die Abkürzungen *p* und *r* stehen für die Evaluationsmaße Präzision und Recall, siehe 3.1.1). Wie Tabelle 3.2 zu entnehmen ist decken die diskutierten Ansätze Sachtexte aus einer Anzahl unterschiedlicher Domänen ab. Mit juristischen Texten befasst sich jedoch keine der Studien. In allen Fällen kommt als Kernkomponente bei der Definitionssuche eine gewisse Zahl von Mustern zum Einsatz, die definitionstypische Formulierungen beschreiben (etwa Kopulakonstruktionen mit modifiziertem Prädikativum: *A is a B which C*). Diese *Definitionsmuster* wurden in allen Ansätzen aufgrund von Korpusbeobachtungen und Introspektion manuell spezifiziert. Die Autoren nutzen größtenteils Suchmuster in Form von Wortfolgen oder Wort-/POS (*part of speech*, Wortart)-Sequenzen. Einige der in Pearson (1998), Del Gaudio und Branco (2007) sowie Westerhout und Monachesi (2007) angegebenen Muster setzen zusätzlich eine Erkennung von Nominal- und Präpositionalphrasen voraus. In mehreren Fällen wird die musterbasierte Suche durch Information aus zusätzlichen Quellen weiter präzisiert: Pearson kombiniert die Definitionsidentifikation mit einer regelbasierten Terminologie-Erkennung und extrahiert nur definitorische Passagen in denen zugleich ein wohlgeformter Term erkannt wurde. Barrière diskutiert zusätzlich zu strukturellen Suchmustern die Nutzung von Information aus *Wordnet* (verwendet werden Synonymmengen und semantische Klassen) im Extraktionsprozess, wobei unklar bleibt, ob diese in dem von ihr diskutierten System tatsächlich eingesetzt wird. Westerhout und Monachesi nutzen automatisch trainierte Klassifizierer zur Filterung der musterbasierten Suchergebnisse, vgl. auch unseren Ansatz in Kap. 6.

| Referenz | Domäne (Sprache) | Suchmuster (Zahl / Analysen) | Zusatzinformation/ Evaluation |
|---|---|---|---|
| Pearson (1998) | Wissenschaftlich-technischer Text (en) | k.A./ Phrasen | Termform-Regeln |
| Meyer (2001) | Populärwissenschaftlicher Text (en) | k.A./ k.A. | Suchmuster aus Lexikondefinitionen |
| Sierra und Alarcón (2002) | Katastrophen-Management (es) | 90/ k.A. | Termform-Regeln |
| Barrière (2004b,a) | Scuba-diving (en) | k.A./ Stämme, POS | keine/ p=0,83[3] |
| Storrer und Wellinghoff (2006) | Texttechnologie (de) | 19/ k.A. | |
| Przepiorkowski u. a. (2007) | Kursmaterial (pl) | 48/ POS | keine/ p=0,19, r=0,6 |
| Del Gaudio und Branco (2007) | Kursmaterial (pt) | k.A./ Phrasen | keine/ p=0,14, r=0,86 |
| Westerhout und Monachesi (2007) | Kursmaterial | 67/ Phrasen | Klassifizierer als Filter[4]/ p=0,8, r=0,67 p=0,5, r=0,36[5] |

Tabelle 3.2: Ansätze zur Identifikation definitorischer Textpassagen

### (b) Evaluation

Für die Evaluation von Definitionsextraktions-Systemen liegen wiederverwendbare Bezugskorpora (wie sie für die etablierten Forschungsbereiche des Informationszugriffs im Rahmen der oben diskutierten *shared task*-Wettbewerbe erstellt wurden) mit annotierten Definitionen bisher nicht vor. Eine besondere Schwierigkeit für die Erstellung solcher Referenzressourcen besteht darin, dass – wie in den ersten beiden Kapiteln dieser Arbeit ausführlicher diskutiert – der Definitionsbegriff nicht ohne weiteres allgemeingültig und domänenunabhängig zu bestimmen ist. Dafür, welche Information in einer Definition angegeben wird, und somit auch dafür, ob eine bestimmte Textpassage als Träger definitori-

[3] Für eine Auswahl von insgesamt sechs Suchmustern, die Definitionen durch funktionale Attribute modellieren. Barrière betrachtet zudem für die von ihr als *productivity* bezeichnete Abweichung von der durchschnittlichen Anzahl an *true positives* pro Muster und nennt Werte zwischen -91% und +127%.

[4] basierend auf verschiedenen Merkmalen wie *Kontextwörter* und *Definitheit*

[5] Kopulabasierte bzw. satzzeichenbasierte Suchmuster und Klassifikation

scher Information anzusehen ist oder nicht, sind je nach Domäne, Blickwinkel, Zielsetzung und Hintergrundwissen unterschiedliche Faktoren relevant. Ebenso variieren die sprachlichen Mittel zur Realisierung von Definitionen zu einem erheblichen Teil domänen- und textsortenspezifisch. Auch Goldstandardkorpora für Definitionseextraktions-Aufgaben können daher kaum domänenübergreifend verwendet werden. Eigene Goldstandards für verschiedene Domänen und Textsorten sind unverzichtbar. Aufgrund der relativ geringen Häufigkeit von Definitionen in den meisten Textsorten ist die Erzeugung solcher Goldstandards jedoch noch erheblich arbeitsintensiver als in anderen Bereichen.

Von den in Tabelle 3.2 aufgeführten Ansätzen verzichten die ersten drei vollständig auf eine quantitative Evaluation.[6] Barrière beschränkt sich bei der Evaluation auf die Berechnung der ohne Goldstandard ermittelbaren Größen Präzision und *productivity* (Abweichung von der durchschnittlichen Anzahl an *true positives* pro Muster). Für deren Ermittlung wählt sie die einfachste mögliche Vorgehensweise: Sie nimmt (unabhängig von Redundanzen und enthaltener Zusatzinformation) eine binäre Klassifikation aller Treffer als [±definitorisch] vor.

Die verbleibenden Ansätze ermitteln auch Recall-Werte und benötigen daher ein Referenzkorpus. Sie nutzen projektspezifisch manuell erzeugte Goldstandards, die jeweils eine für einen einzelnen Annotator überschaubare Anzahl von Dokumenten (< 30) umfassen. Nähere Angaben zur Annotation (etwa zur Anzahl der Annotatoren oder ggf. zum Inter-Annotator Agreement) werden in keiner der zitierten Studien gemacht. Die in Tabelle 3.2 angegebenen Evaluationsergebnisse sind daher nur als ungefähre Orientierung zu bewerten und ermöglichen keine aussagekräftigen Vergleiche.

### 3.2.2 Verarbeitung von Definitionen zu strukturierten Ressourcen

Die Nutzbarmachung textueller Definitionen als Quelle für die Erzeugung strukturierter Wissensressourcen ist bereits seit relativ langer Zeit immer wieder Gegenstand computerlinguistischer Forschungen gewesen. Die Verarbeitung von Definitionen in freiem Text zu einer strukturierten Ressource erfordert die Identifikation und die strukturelle Analyse von Definitionen sowie abschließend deren Integration in eine Wissensstruktur. Die ersten beiden Schrit-

[6] Neben den angesprochenen grundsätzlichen Problemen mag ein Grund hierfür das Anwendungsszenario sein, auf das diese Ansätze abzielen. Bei der praktischen Nutzung in der Terminologiearbeit würden die Ergebnisse der betrachteten Systeme zur Durchsicht, Auswahl und weiteren Aufarbeitung in Form einer KWIC (*keyword in context*)-Liste an einen Experten weitergegeben. Entscheidend ist für diesen Zweck vor allem, dass eine gewisse Anzahl repräsentativer Fundstellen identifiziert wird, nicht dass alle definitorischen Passagen ermittelt oder ausschließlich *true positives* geliefert werden.

te können als eine anspruchsvolle Form der IE betrachtet werden, bei der eine vorgegebene "Begriffsschablone" mit definitorischen Wissensbestandteilen aus einer Textbasis zu füllen ist. Allerdings werden auf sprachlicher Ebene in Definitionen nicht (wie etwa in den Relationen und Events der MUC-Szenarios) allein *Named Entities* oder andere kompakte nominale Ausdrücke verknüpft. Sowohl das Definiens als auch der definierte Terminus selber können durch komplexe (mitunter satzwertige) deskriptive Ausdrücke realisiert werden. Die hierarchische Architektur der meisten IE-Systeme (vgl. 3.1.2) kann daher auf die Extraktion und Verarbeitung von Definitionen nicht ohne weiteres übertragen werden.

Ein Großteil der bisherigen Forschungen hat sich deshalb auf die Umsetzung von Definitionen aus bereits hochstrukturiertem und einheitlich formatiertem Text, vornehmlich aus Lexikon- oder Glossareinträgen, beschränkt und konnte somit den dritten Schritt – Integration in eine Wissensstruktur – fokussieren. Zur Verarbeitung von Definitionen aus freiem Text sind in erster Linie Einzelaspekte untersucht worden, nur wenige Ansätze wurden implementiert und evaluiert.

#### (a) Verarbeitung von Glossar- und Lexikondefinitionen

**Maschinenlesbare Lexika.** Ab Mitte der 1980er Jahre wurden konventionelle Sprachlexika zunehmend bis zur Druckvorstufe in maschinenlesbarer Form erstellt, und es begannen verschiedene Untersuchungen zu Methoden, um das sprachliche Wissen in solchen Lexika automatisch zu erschließen. Die dabei verfolgten Zielsetzungen waren teils allgemein-theoretischer Natur, wie etwa die empirische Identifikation semantischer Primitive durch den Vergleich einer Vielzahl von Lexikondefinitionen, vgl. Calzolari (1984). In den meisten Ansätzen – u.a. innerhalb des europäischen AQUILEX-Projekts – ging es jedoch darum, auf der Basis bestehender Ressourcen Lexikonkomponenten für sprachtechnologische Anwendungen zu schaffen oder zu ergänzen. Vgl. hierzu u.a. Markowitz u. a. (1986); Ravin (1990); Briscoe u. a. (1990, 1993). Für Gesamtdarstellungen siehe Briscoe (1989) und Wilks u. a. (1995).

In den meisten Lexika entsprechen Definitionen einzelnen Einträgen. Sowohl Definitionen als solche als auch deren Grobstruktur (d.h. zumindest die Einteilung in das definierte Stichwort und die zu diesem angegebene Information) sind also aufgrund typographischer oder eindeutiger struktureller Merkmale zu erkennen. Die Aufgabe der Definitionsextraktion kann aus diesem Grund meist mit relativ einfachen Mitteln gelöst werden, Definitionsmuster können unkompliziert als Varianten einiger weniger Grundmuster beschrieben werden, vgl. Alshawi (1987); Ahlswede und Evens (1988); Barnbrook (2002). Neff und

Boguraev (1989) diskutieren jedoch auch eine Vielzahl praktischer Probleme, die sich aus den Formatierungskonventionen von Lexika ergeben können.

Der Schwerpunkt der Forschungen zur Verarbeitung von Lexika lag daher nicht auf der Spezifikation von Extraktionsregeln, sondern auf der genauen semantische Analyse der einzelnen Lexikoneinträge und auf Fragen der Repräsentation und Nutzbarkeit des extrahierten Wissens. Auch die tiefere Eintragsanalyse wird in Lexika durch Formatierungs- und Strukturierungskonventionen (von einheitlichen Lesartmarkierungen bis hin zu in einigen Lexika genutztem kontrolliertem primitivem Vokabular[7]) erleichtert. Jedoch musste vielfach erst inventarisiert werden, welche Information in verschiedenen Lexika explizit enthalten ist und welche (wie etwa die genaue Beziehung zu Oberbegriffen) nur inferiert werden kann (Calzolari (1984); Boguraev und Briscoe (1989); Briscoe u. a. (1990); Vossen und Copestake (1993)). Forschungen zu Repräsentationsfragen waren meist zugleich durch generellere Fragestellungen aus dem Bereich der lexikalischen Semantik motiviert, etwa zur Nutzung von Default-Unifikation für die Modellierung lexikalischer Vererbungsbeziehungen (Russell u. a. (1993); Krieger und Nerbonne (1993)) oder zur Spezifikation lexikalischer Regeln und ähnlicher Mechanismen für die Darstellung generativer Prozesse innerhalb des Lexikons (Boguraev und Pustejovsky (1990); Pustejovsky (1995); Briscoe und Copestake (1999)). Neuere Untersuchungen befassen sich vor allem mit Methoden zur Kombination redundanter Quellen, um Information zu einzelnen Begriffen zu ermitteln (es kommen Text Mining-Verfahren wie Clustering und Alignment zum Einsatz, Sierra und McNaught (2000); Scott Piao und Ananiadou (2008)).

**Nutzung von Online-Ressourcen.** Mit der allgemeinen Verfügbarkeit großer online-Datenbestände seit Mitte der 1990er Jahre wurde versucht, auch diese in ähnlicher Weise wie im vorigen Abschnitt dargestellt zur Erstellung lexikalischer Ressourcen zu nutzen. Zwar sind Definitionen auch in online verfügbaren Lexika, Enzyklopädien und Glossaren meist durch relativ eindeutige Formatierungsmerkmale ausgezeichnet (die zudem im HTML-Format oft sehr viel leichter automatisch zu erkennen sind als in proprietären DTP-Formaten). Jedoch ist ein großer Teil der online verfügbaren Ressourcen nicht so umfangreich und informativ und nicht im gleichen Maße redaktionell aufgearbeitet und nach einheitlichen Konventionen strukturiert wie konventionelle Lexika (vgl. hierzu die Beispiele in Klavans und Whitman (2001)). Fokus der Forschung ist in diesem Kontext deshalb generell die Untersuchung von Verfahren zur Kom-

[7] Vgl. MacFarquhar und Richards (1983). Besondere Beachtung hat das auf etwa 2000 Wörter reduzierte *defining vocabulary* des *Longman Dictionary of Contemporary English* (LDOCE) erfahren, siehe hierzu Summers (1987).

bination partieller und teilweise überlappender Definitionen aus verschiedenen Quellen. Dabei wird u.a. an Methoden zum automatischen Zusammenfassen angeknüpft (Fujii und Ishikawa (2004); Hovy u. a. (2003)). Ein Beispiel für die Erschließung von online-Definitonsbeständen in einem Retrieval-System stellt die im Webangebot von Google verfügbare *define*-Funktion dar, durch die eine Suche auf Glossar-Webseiten eingeschränkt wird.

#### (b) Verarbeitung von Definitionen in freiem Text

Die meisten Untersuchungen zur Verarbeitung von Definitionen in freiem Text befassen sich entweder mit konzeptuellen Fragen, z.B. den Eigenschaften, die Begriffsnetzwerke für die Repräsentation der in Definitionen enthaltenen Information benötigen (vgl. Büchel und Weber (1995); Barrière (2004a)), oder betrachten Einzelaspekte des Analyseprozesses wie die Identifikation und themenzentrierte Gruppierung von Domänenterminologie (Park u. a. (2002); Liu u. a. (2003)) oder die Erkennung bestimmter Relationstypen, etwa *is-a* oder *part-of*, in Definitionen (Malaisé u. a. (2004); Malaise u. a. (2005)).

Muresan u. a. (2003) beschreiben das unseres Wissens bisher einzige implementierte System, das Definitionen aus freiem, nicht vorstrukturiertem Text in eine strukturierte Ressource umsetzt. Es nutzt den musterbasierten *Definder*-Ansatz (Klavans und Muresan (2001a,b, 2002)). Dieser wurde zunächst zu Erzeugung von Glossaren aus populärwissenschaftlichen und Fachtexten in der biomedizinischen Domäne entwickelt. Dann wurde er um nicht näher beschriebene regelbasierte und statistische Methoden zur Erkennung von Attributen und semantischen Relationen in Definitionen sowie um eine Komponente zum Transfer in eine Datenbank ergänzt. Die Autoren nennen Präzisions- und Recall-Werte von 0,86 bzw. 0,84 für die Definitionsextraktion und einen Gesamtzahl von 12 780 Definitionen für 8431 Begriffe in der aus 147 (teilweise Glossare enthaltenden) Webseiten und 600 weiteren Dokumenten erzeugten Datenbank. Sie gehen jedoch nicht näher auf die Evaluationsmethodologie und -modalitäten ein, so dass die Aussagekraft dieser Ergebnisse nur schwer einzuschätzen ist.

### 3.2.3 Definitionsfragen im Question Answering

Eine weitere relativ weitreichende Aufgabenstellung im Zusammenhang mit der textbasierten Identifikation von Definitionen ist die Beantwortung von Definitionsfragen. Bei Definitionsfragen handelt es sich um Fragen der Form *Was / Wer ist... ?*, mit denen (zumindest in einer in vielen Zusammenhängen sinnvollen Lesart) eine Bestimmung des in der Frage spezifizierten Begriffs erfragt wird. Definitionsfragen spielten besonders im Rahmen von TREC 12 (vgl.

Voorhees (2003)) eine bedeutende Rolle und wurden dort erstmals systematisch evaluiert. Der Fragenkatalog dieses Wettbewerbs enthielt fünfzig Definitionsfragen, dreißig davon *biographische Fragen* nach Personen (z.B. *Ben Hur*), zehn Fragen nach Organisationen (z.B. *Bausch & Lomb*) und zehn nach sonstigen Konzepten (wie *Feng Shui*). Der gesuchte Informationstyp wurde durch ein Szenario umschrieben: Ein durchschnittlich gebildeter Zeitungsleser trifft auf einen Begriff, zu dem er nähere Informationen erhalten möchte.

Ab TREC 13 wurden alle Fragen um Begriffe gruppiert. Dabei entsprachen jeweils die letzten Fragen zu einem Begriff den bisherigen Definitionsfragen, jedoch waren sowohl die Bezeichnung (*other questions*) als auch die Anweisung zur Beantwortung (*tell me other interesting things about X*) deutlich allgemeiner gehalten als in TREC 12 (vgl. Voorhees (2004)). Der CLEF-Fragenkatalog enthielt erstmals im Jahr 2004 etwa zwanzig Definitionsfragen für jede untersuchte Sprache, wobei bewusst auf Fragen nach Definitionen allgemeiner Konzepte verzichtet wurde (Magnini u. a. (2004)).

#### (a) Besonderheiten

Definitionsfragen unterscheiden sich in drei wichtigen Punkten von faktoiden Fragen, für deren Beantwortung ein großer Teil der bisher entwickelten QA-Systeme optimiert wurde:

- Sie sind in der Regel nicht durch ein einzelnes Wort oder eine subsententielle Phrase zu beantworten, sondern erfordern komplexere, satzwertige oder sogar mehrsätzige Antworten.
- Sie lassen mehr Antwortvarianten zu als faktoide Fragen. Insbesondere sind häufig viele und inhaltlich unterschiedliche partielle Antworten möglich.
- Der Frageformulierung ist – bis auf den Zielbegriff, zu dem eine Definition erfragt wird – keine weitere Information über mögliche Antworten zu entnehmen. Unter anderem bedeutet dies, dass für Definitionsfragen das im QA ansonsten übliche Verfahren, anhand der Frage auf einen erwarteten Antworttyp zu schließen und dieses Wissen zur Steuerung der Antwortsuche zu nutzen, nicht anwendbar ist.

Zugleich besteht in der Tatsache, dass Definitionsfragen einen Zielbegriff spezifizieren, ein wesentlicher Unterschied zur Aufgabe der Definitionsextraktion, mit der sich die in den vorigen Abschnitten vorgestellten Ansätze befassen: Bei der Suche nach Antworten auf Definitionsfragen kann ebenso wie bei der Evaluation solcher Antworten *begriffszentriert* vorgegangen werden (vgl.

Storrer und Wellinghoff (2006)). Für die Antwortextraktion bedeutet dies zunächst, dass begriffsspezifische Information zur Fokussierung der Suche verwendet werden kann. Weiterhin müssen nicht zwingend alle Auftreten semantisch gleichwertiger definitorischer Information extrahiert werden. Redundanzen in der Textgrundlage (oder auch zusätzliche redundante Quellen) können vielmehr zur Verbesserung der Ergebnisqualität genutzt werden.

#### (b) Ansätze

Die meisten Ansätze zum QA mit Definitionsfragen wurden im Zusammenhang mit dem TREC 12-Wettbewerb entwickelt und sind daher auf die Antwortsuche in englischsprachigem Zeitungstext spezialisiert (das bei den TREC-Wettbewerben verwendete Aquaint-Korpus umfasst englischsprachige Nachrichtenmeldungen aus dem *Xinhua*-, *New York Times*- und *Associated Press Worldstream*-News Service, vgl. Graff (2002)). Der Entwicklung spezialisierter Definitionsfragen-Systeme für andere Sprachen und Textsorten ist bisher vergleichsweise wenig Aufmerksamkeit zugekommen. Auf nicht-englischen Text stützen sich unseres Wissens nur die im Rahmen der CLEF-QA-Tracks evaluierten Systeme der Universitäten von Groningen (für das Niederländische) und des mexikanischen INAOE-UPV (für das Spanische, Denicia-Carral u. a. (2006)), der auf der Grundlage des TREC 13-Systems von BBN Technologies entwickelte Ansatz für das Chinesische (Peng u. a. (2005)) sowie das Web-QA / Mdef-WQA-System für multilinguales WWW-basiertes QA, beschrieben in Figueroa und Neumann (2007), Figueroa (2008b) und Figueroa (2008a). Yu u. a. (2007) beschreiben mit MedQA ein ohne Bezug zu QA-Wettbewerben entwickeltes System, das auf die Textsorte *Medizinischer Fachtext* zugeschnitten ist. Als Textgrundlage dient unter anderem eine Sammlung von 15 Millionen MEDLINE-Einträgen. In Anhang B sind einige Charakteristika der wichtigsten bisher implementierten Systeme zum QA mit Definitionsfragen sowie einiger Ansätze zusammengestellt, die sich mit Einzelkomponenten innerhalb einer gesamten Architektur befassen.

QA-Systeme für Definitionsfragen stützen sich im allgemeinen auf die in Abb. 3.3 dargestellte generische QA-Architektur, die sie jedoch um Filterungsphasen im Rahmen der Antworterkennung ergänzen. In den meisten Fällen kommt zunächst ein Filter zum Einsatz, das auf sprachlichen *Definitionsmustern* basiert, wie sie auch von den im vorigen Abschnitt besprochenen Definitionsextraktionssystemen genutzt werden.[8] Bis auf die Ansätze von Harabagiu

[8] Vefahren, die völlig auf die Nutzung von Definitionsmustern verzichten, sind in Prager u. a. (2001), Ahn u. a. (2004) und Zhang u. a. (2005) beschrieben. Stattdessen werden Antwortprojektionstechniken bzw. bestimmte Wordnet-Hypernyme des Suchbegriffs zur Definitionsidentifikation genutzt. Der letztgenannte Ansatz basiert auf der Intuition, dass Begriffe meistens

u. a. (2003) und das WebQA / Mdef-WQA-System verwenden alle in Anhang B aufgeführten Systeme zusätzlich zur musterbasierten Filterung noch eine zweite Filterungsphase, in der die Zahl möglicher Antwortpassagen aufgrund von Zusatzinformation weiter eingeschränkt wird.

**Definitionsmuster.** Art und Anzahl der in den verschiedenen Ansätzen verwendeten Definitionsmuster unterscheiden sich deutlich. In vielen Systemen basieren die Muster auf Wort- bzw. Wort-/Wortartsequenzen. Teilweise werden jedoch auch partielle syntaktische Strukturen verwendet (Blair-Goldensohn u. a. (2003a,c,b)). Die von Echihabi u. a. (2003) genutzten Definitionsmuster arbeiten auf der Ebene semantischer Relationen. Während einige Ansätze (z.B. Gaizauskas u. a. (2003)) viele und stark optimierte Suchmuster nutzen, stützen sich andere (etwa Han u. a. (2005, 2006a,b) und Xu u. a. (2005, 2006)) auf eine geringe Zahl ergiebiger Muster und setzten somit stärker auf die Ausnutzung von Redundanzen in der Textbasis.

Der Musterabgleich erfolgt in aller Regel durch strikte Substring-Suche bzw. reguläre Ausdrücke. Eine Ausnahme bildet hier das in Cui u. a. (2004a, 2005) beschriebene System, das einen probabilistischen Musterabgleich-Prozess (sog. *soft pattern matching*) verwendet, der partielle und nicht-kontingente Überlappungen zwischen Antwortkandidat und Definitionsmuster zulässt.

In fast allen Fällen sind die genutzten Definitionsmuster aufgrund von korpusbasierten Pilotstudien, Einzelbeobachtungen und Introspektion ausgewählt und manuell spezifiziert worden. Nur das CLEF-System des mexikanischen INAOE-UPV sowie der in Cui u. a. (2004a, 2005) diskutierte Ansatz arbeiten mit Definitionsmustern (in Form von Wortsequenzen), die durch Bootstrapping-Verfahren automatisch akquiriert wurden.

**Filterung auf der Basis von Zusatzinformation.** Auch die zusätzliche Filterung der Antwortkandidaten beruht in den verschiedenen Ansätzen auf teils deutlich unterschiedlichen Arten von Information. Es kommen zum einen (a) *datengestützte Definitionsprofile* zum Einsatz, zum anderen (b) verschiedene *heuristisch motivierte Merkmale*.

(a) *Definitionsprofile* repräsentieren die Lexik von Definitionen in einem Referenzbestand. Als solcher dienen meistens frei verfügbare Definitionssammlungen (Wordnet-*glosses*, Webseiten mit Enzyklopädieartikeln oder Biographien). Das DefScriber-System (Blair-Goldensohn u. a. (2003a,c,b)) nutzt zudem als Fallback die Menge aller extrahierten Definitionskandidaten.

unter Verwendung von Oberbegriffen mittlerer Spezifizität definiert werden (entsprechend den in Rosch u. a. (1976) postulierten *basic objects*).

Es lassen sich drei verschiedene Typen von Definitionsprofilen unterscheiden:

1. *Allgemeine Definitionsprofile.* Diese beinhalten lexikalische Elemente, die in einem Referenzbestand von allgemeinen Definitionen besonders häufig auftreten.

2. *Definitionstypenprofile* Sie stützen sich auf die Lexik der Definitionen eines bestimmten Begriffstyps (also z.B. *Person* oder *Organisation*)

3. *Begriffsprofile* enthalten lexikalische Elemente, die mit dem Zielbegriff – allgemein oder in Definitionen – besonders häufig kookurrieren oder in Definitionen des Zielbegriffs in externen Ressourcen gefunden wurden.

Die jeweils verwendeten Profile werden in den verschiedenen Ansätzen auf im einzelnen unterschiedliche Weise repräsentiert. Die Filterung erfolgt jedoch in den meisten Fällen durch einen vektorbasierten Vergleich zwischen Definitionsprofil und extrahierten Definitionskandidaten. In manchen Ansätzen bilden Profile zudem eine Quelle für zusätzliche Suchbegriffe in der IR-Phase. Chen u. a. (2006) und Han u. a. (2006b) repräsentieren ein Begriffs- und ein allgemeines Definitionsprofil als Sprachmodelle. Sie nutzen dann die Abschätzung der jeweiligen Auftretenswahrscheinlichkeiten durch diese Modelle als Grundlage für ein Ranking der Definitionskandidaten.

(b) *Heuristisch motivierte Merkmale* werden in den meisten Systemen zur Erzeugung einer Rangfolge von Definitionskandidaten genutzt, Fahmi und Bouma (2006) implementieren eine klassifikationsbasierte Filterung anhand heuristischer Merkmale, vgl. auch den in 3.2 genannten Ansatz von Westerhout und Monachesi (2007).

Zu den verwendeten Merkmalen gehören u.a.:

- Zuverlässigkeit der Suchmuster auf einem Testkorpus
- Position extrahierter Sätze im Quelldokument
- Hypero- / Hyponymrelation zwischen Begriffen in einem Definitionskandidaten
- Ausgewählte Relationen und Propositionen im Definitionskandidaten (die durch ein IE-System ermittelt werden)

Androutsopoulos und Galanis (2005) und Miliaraki und Androutsopoulos (2004) untersuchen systematisch die Effektivität verschiedener Merkmalskombinationen.

Prager u. a. (2003) implementieren zusätzlich eine von ihnen als *QA by dossier* bezeichnete heuristische Systemkomponente: Auf der Basis des Zielbegriffs der Definitionsfrage werden nach heuristischen Regeln faktoide Fragen nach Teilaspekten erzeugt (also beispielsweise nach Geburtsort und -datum einer Person) und durch andere QA-Komponenten beantwortet.

**Antwortgenerierung.** Die in TREC 13 für Systemantworten auf Definitionsfragen erwartete Form war eine ungeordnete Liste, bestehend aus Textausschnitten und Verweisen auf das jeweilige Quelldokument. Weder für die Gesamtliste noch für die einzelnen Textausschnitte war eine Maximallänge vorgegeben, allerdings bevorzugte das zur Evaluation verwendet Maß (s.u.) kurze Antworten. Eine kohärente Anordnung sowie kohäsive Verknüpfung der einzelnen Listeneinträge war nicht erforderlich.

Entsprechend liefert – bis auf DefScriber (Blair-Goldensohn u. a. (2003a,c,b)) – keines der in Anhang B aufgelisteten Systeme eine auf die Erfüllung von Textualitätskriterien ausgerichtete Antwort. Um eine Abwertung aufgrund zu großer Antwortlängen zu vermeiden, versuchen die meisten Systeme jedoch, redundante und irrelevante Ergebnisse aus der endgültigen Antwort auszuschließen. Dabei kommen verschiedene, der Forschung zur automatischen Textzusammenfassung entnommene Techniken zum Einsatz, beispielsweise:

- Clustering und Auswahl von Repräsentanten der einzelnen Cluster
- *Minimal Marginal Relevance* (MMR)-Verfahren, bei denen im Anschluss an ein Ranking nur solche Textausschnitte beibehalten werden, die sich hinreichend stark von höherrangigen Ergebnissen unterscheiden (Carbonell und Goldstein (1998))
- Auswahl häufiger gemeinsamer Bestandteile einzelner Ergebnisse (bestimmt nach unterschiedlichen heuristischen Überlappungsmaßen)

Blair-Goldensohn u. a. (2003a) nutzen eine Kombination mehrere solcher Techniken sowie eine hohe Gewichtung einzelner Definitionsmuster, die sie als besonders informativ ansehen, um zusätzlich zur Satzauswahl eine textuell akzeptable Satzreihenfolge in ihrer Systemantwort sicherzustellen.

### (c) Evaluation

Aufgrund der begriffszentrierten Anlage der Aufgabenstellung kann bei der Evaluation definitorischer QA-Systeme auf die Erstellung eines Goldstandards

verzichtet werden. Stattdessen wird – ausgehend von den Zielbegriffen – introspektiv die Menge der für eine Definition benötigten Informationsbestandteile bestimmt und dann das Vorhandensein jeweils eines Auftretens im Textkorpus verifiziert. Auf diese Weise werden schon bei der Erstellung für jede Trainings- und Testfrage erwartete Antwortelemente zusammengestellt. So erstellte *Antwortschlüssel* bildeten sowohl in den TREC-Wettbewerben als auch im Rahmen der CLEF-Wettbewerbe die Grundlage der Evaluation. Die TREC-Antwortschlüssel umfassten für eine Frage in der Regel mehrere als *vital* und zusätzlich mehrere als *non-vital* klassifizierte Informationseinheiten. Alle Auftreten dieser sog. *nuggets* in den Systemantworten wurden dann markiert. Zur Bestimmung des Recall eines Systems wurde ermittelt, wie viele der *vital nuggets* in dessen Antworten enthalten waren. Die Präzision bestimmte sich (da die Bestimmung der Gesamtzahl an *nuggets* in einer beliebigen Antwort als prinzipiell nicht durchführbar angesehen wurde) über die Länge der Antworten.

In den CLEF-Wettbewerben wurde hingegen schon bei der Zusammenstellung des Fragekatalogs darauf geachtet, dass für alle Definitionsfragen möglichst einfache, auf Einzelfakten beruhende Antworten existieren (vgl. Magnini u. a. (2004), 374 ff.), so dass eine direkte binäre Klassifikation der Systemantworten als richtig oder falsch möglich war.

Die beschriebene Vorgehensweise stützte sich in beiden Wettbewerben auf menschliche Urteile nicht nur bei der Erzeugung der Antwortschlüssel, sondern auch bei der Bewertung der einzelnen Systemantworten. Zur Vermeidung dieser Abhängigkeit experimentieren verschiedene Autoren mit Varianten des ROUGE-Maßes (das ursprünglich zur Bewertung automatischer Zusammenfassungen eingesetzt wurde, vgl. Lin und Hovy (2003)) beim Abgleich zwischen Systemantworten und Antwortschlüsseln (Xu u. a. (2004); Peng u. a. (2005); Cui u. a. (2005); Han u. a. (2005); Lin und Demner-Fushman (2005)). Xu u. a. (2004) stellen eine starke Korrelation mit den Ergebnissen einer vollständig manuellen Evaluation fest.

## 3.3 Informationstechnologie für juristische Anwendungen

Im Rest dieses Kapitels geben wir einen Überblick über die Nutzung von Informationstechnologie in der juristischen Domäne. Wir gehen dabei zunächst auf fortgeschrittene informationstechnologische Lösungen für juristische Anwendungen im allgemeinen ein und kommen dann speziell auf die Rolle von Systemen für den textbasierten Informationszugriff zu sprechen. Da die in diesem Teil des Kapitels relevanten Ansätze in der Literatur größtenteils nicht so

detailliert beschrieben sind wie die bisher besprochenen Verfahren, werden wir auch in unseren Erläuterungen kaum auf ihre genaue technische Umsetzung eingehen können.

Elektronische Entscheidungssammlungen stellen inzwischen in der juristischen Praxis ein unverzichtbares Hilfsmittel dar. Die zur Erschließung dieser Information verwendeten Systeme bleiben jedoch hinter dem in den vorigen Abschnitten skizzierten Stand der Technik zum automatisierten Informationszugriff zurück. Auch im akademischen Bereich sind bisher nur wenige Versuche zur Verwendung moderner Infomationszugriffs-Techniken in der juristischen Domäne erfolgt (obwohl die Nutzung fortgeschrittener Informationstechnologie, insbesondere von "KI-Techniken", für juristische Anwendungen durchaus regelmäßig Gegenstand von Forschungsprojekten gewesen ist).

Für diese Situation dürften zunächst einmal historische Gründe verantwortlich sein: Das Augenmerk der Forschungen im Berührungsfeld von Informatik und Recht (wir gebrauchen in diesem Sinne auch den in der deutschen Literatur sehr uneinheitlich verwendeten Begriff *Rechtsinformatik*) galt lange Zeit in erster Linie der Nachbildung des richterlichen Entscheidungsprozesses durch die Modellierung juristischen Wissens, Schlussfolgerns und Argumentierens. Bereits 1949 machte Lee Loevinger den Vorschlag, Computer zur Lösung von Rechtsproblemen zu verwenden (Loevinger (1949)), und eine zentrale technische Zielsetzung der Rechtsinformatik blieb im Anschluss die Umsetzung entscheidungsunterstützender Expertensysteme. Das mit den Entwicklungen im Bereich des Informationszugriffs und der Sprachtechnologie gegebene Potential zur Unterstützung beim Zugang zu juristischen Wissensbeständen wurde hingegen relativ spät erkannt und ist erst in den letzten Jahren stärker fokussiert worden.

Im Folgenden gehen wir zunächst anhand einiger Beispiele auf diese historische Entwicklung des rechtsinformatischen Forschungsinteresses ein (3.3.1-3.3.4). Wir schließen das Kapitel dann (in 3.4) mit der Diskussion einiger prinzipieller Schwierigkeiten ab, die sich für juristische Informationszugriffs-Systeme aufgrund von Besonderheiten der Rechtsdomäne – also unabhängig von historischen Gründen – ergeben.

### 3.3.1 Expertensysteme

In den 1970er und 1980er Jahren wurde eine Vielzahl juristischer Expertensysteme für unterschiedlichste Aufgaben entwickelt (Jandach (1993) zählt 119). Die dabei verfolgten Ansätze reichen von der direkten Beschreibung von Tatbestand-Rechtsfolge-Paaren durch Produktionsregeln (etwa im *Latent Damage Advisor* System, Capper und Susskind (1988)) über die Formalisierung kompletter Regelungszusammenhänge in Form von Horn-Klauseln (wie in der in

Sergot u. a. (1986) beschriebenen Formalisierung des britischen Staatsangehörigkeitsrechts), semantischen Netzen oder Wissensrahmen (so im *TAXMAN*-System, McCarty (1980, 1977)) bis zur Nutzung rechtstheoretisch fundierter Spezialformalismen (z.B. McCartys *language for legal discourse*, McCarty (1989)).

Auch wenn einige juristische Expertensysteme – vor allem in der öffentlichen Verwaltung, zum Teil auch von Anwälten – erfolgreich eingesetzt wurden und werden, hat ein Großteil der verfolgten Ansätze nicht zu praxisreifen Anwendungen geführt. Insgesamt werden die entsprechenden Bemühungen aus heutiger Sicht, zumindest gemessen an den teils hoch gesteckten Zielen, als wenig erfolgreich bewertet. In Deutschland gilt dies insbesondere für das LEX-Projekt (Haft und Lehmann (1989)), das im Bereich der juristischen Expertensysteme Mitte der 1980er Jahre zunächst große Beachtung fand. Die Zielsetzung dieser Kooperation des Wissenschaftlichen Zentrums der IBM Deutschland und der Universität Tübingen, nämlich die automatische Beurteilung unfallrechtlicher Fälle im natürlichsprachlichen Dialog mit einem Benutzer, erwies sich als so ehrgeizig, dass die wichtigsten Projektergebnisse schlussendlich vor allen Dingen in grundlegenden Erkenntnissen über die Grenzen der verwendeten Methodik lagen.

Nicht zuletzt aufgrund dieser Erfahrungen beschränken sich neuere Forschungen zu juristischen Expertensystemen meist auf kleinere Domänen und die Untersuchung konzeptueller Alternativen für die Umsetzung von Einzelkomponenten in bestehenden generischen Architekturen. So wurden in jüngerer Zeit Methoden zum Umgang mit unscharfem Wissen sowie zur dynamischen Integration unterschiedlicher Wissensquellen erprobt. Das 1998 abgeschlossene DFG-Projekt *Fuzzy-Schmerzensgeld* an der Universität des Saarlandes etwa untersuchte die Modellierung des Entscheidungsprozesses bei der Schmerzensgeldbemessung unter Anwendung fuzzy-logischer Methoden (Zadeh (1974, 1965)). Bohrer (1993) beschreibt ein Expertensystem zu urheberrechtlichen Fragen, das dynamisch online verfügbare Wissensbestände in den Prüfungsprozess integrieren kann.

### 3.3.2 Ontologien

Ab Mitte der 1990er Jahre widmete sich die rechtsinformatische Forschung auch verstärkt anderen Themen als der Entwicklung von Expertensystemen. Ein wichtiges solches Thema ist die Repräsentation juristischen Wissens. Dabei wurden zunächst vor allem die konzeptuellen Grundlagen von Rechtsontologien (also formal repräsentierter juristischer Begriffsnetzwerke) untersucht. So schlagen Valente und Breuker (1994) und Valente (1995) funktional bestimmte Wissenstypen (z.B. normatives, reaktives oder kreatives Wissen) als Grund-

kategorien einer juristischen Ontologie vor. Dagegen verwenden van Kralingen (1995) sowie Visser und Bench-Capon (1996) Normen, Handlungen und Konzepte als Primitive und sehen für diese eine Repräsentation in Form von Wissensrahmen als angemessen an. In der Folge wurden zum einen juristische *upper level*-Modelle entworfen (Breuker und Hoekstra (2004); Hoekstra u. a. (2007)), also Modelle, die allgemeine, Rechtsbereichs-übergreifend relevante Begrifflichkeiten und ihre Relationen enthalten. Zum anderen wurden verschiedene Domänenontologien zu speziellen Bereichen einiger Rechtssysteme entwickelt (etwa die CLIME-Ontologie des niederländischen Schifffahrtsrechts, Boer u. a. (2001), oder die im Rahmen des europäischen FF POIROT-Projekts entwickelte *financial fraud ontology*, Leary u. a. (2003); Kingston u. a. (2004)). Ein neuerer Forschungsbereich ist der Aufbau multilingualer Ressourcen (wobei gleichzeitig theoretische Probleme des Rechtsvergleichs zu lösen sind, vgl. Schäfer (2005); Vandenberghe u. a. (2003)).

Ergebnis der angesprochenen Bemühungen waren bisher relativ kleine, durch Experten handkodierte Ressourcen, die unseres Wissen noch nicht in Anwendungen außerhalb des akademischen Kontextes eingesetzt worden sind.

### 3.3.3 Dokumentenmanagement

Als ein bedeutendes praktisches Anwendungsfeld rechtsinformatischer Lösungen trat dagegen in den 1990er Jahren das juristische Informations- und Dokumentenmanagement hervor. Zentrale Bereiche sind dabei (neben der inzwischen durch verschiedene kommerziell verfügbare Werkzeuge unterstützen intelligenten Verwaltung von Fallakten und ähnlichen Dokumenten) der Entwurf von Gesetzen und Verordnungen sowie die Verwaltung von Normtexten und Beweismaterialien zum Zugriff während des Prozesses.

Im Bereich des sog. *legislative drafting* existieren zum einen Systeme, die von kommerziellen Anbietern im Auftrag der öffentlichen Verwaltung für einzelne Aspekte der täglichen Arbeit entwickelt wurden. Sie unterstützen beispielsweise die Einhaltung rechtsförmlicher und redaktioneller Vorgaben oder nehmen Strukturprüfungen vor (etwa die von der Firma Dialogika für die Europäische Kommission bzw. das BMJ implementierten Programme LegisWrite und eNorm). Zum anderen wurden im Rahmen verschiedener Forschungsprojekte anwendungsnahe Prototypen für allgemeinere Aufgabenstellungen, z.B. das Lebenszyklus- und Versionsmanagment von Gesetzestexten (Boer u. a. (2004)) oder die graphische Modellierung von Regelungszusammenhängen und ihre Abbildung auf institutionelle Abläufe (Boer u. a. (2003)) erstellt. Schließlich existieren eine Vielzahl von Initiativen zur Entwicklung XML-basierter Formate für die Gesetzgebung, ohne dass sich jedoch ein übergreifender Standard durchgesetzt hätte. Ein aktueller Bericht des vom *United Na-*

*tions Department of Economic and Social Affairs* und der *Inter-parliamentary Union* gegründeten *Global Centre for Information and Communication Technologies in Parliament* (Biasiotti u. a. (2008)) zählt allein in Europa zehn in diesem Zusammenhang relevante, teils auf nationaler Ebene angesiedelte und teils internationale, Projekte auf.

### 3.3.4 Textbasierter Informationszugriff

Zur Suche in den elektronischen Entscheidungssammlungen kommerzieller Anbieter stehen in der Regel konventionelle Suchmaschinen auf der Basis von Volltextindizes und Bool'schen Verknüpfungen zur Verfügung, die zusätzlich eine Einschränkung nach bibliographischen und juristischen Metadaten (etwa Datum und Gericht) erlauben. Teilweise werden zusätzlich manuell erstellte Schlagwortverzeichnisse und eine Navigation über Zitate im Text angeboten. Fortgeschrittene Verfahren zum Informationszugriff spielen bei kommerziellen Diensten bisher nur eine geringe Rolle, werden jedoch in einzelnen akademischen Forschungsprojekten erprobt.

Zunächst galt das Interesse dabei hauptsächlich der Nutzung manuell kodierter Ressourcen zur Anfrageerweiterung im IR, um eine konzeptuelle Suche anstatt eines rein volltextbasierten Zugriffs zu ermöglichen (Hafner (1978); Bing (1987b,a)). In der Folge wurden zusätzlich Ranking-Verfahren (De Mulder und van Noortwijk (1994)) sowie alternative Retrievalmodelle (etwa das Vektorraummodell in Smith u. a. (1995); Smith (1997) bzw. – mit semantischer Generalisierung durch einen Thesaurus – in Schweighofer und Winiwarter (1995)) eingesetzt. Außerdem wurden Methoden zur Identifikation relevanter Dokumente anhand von Ähnlichkeiten auf der Fallebene erprobt (Rissland und Daniels (1995); Brüninghaus und Ashley (2005)). Inzwischen versuchen einige Ansätze multilinguale Retrieval-Aufgaben mit Ressourcen zu lösen, die sich auf die multilinguale Wordnet-Erweiterung EuroWordnet, Vossen (1998), stützen (so zum Beispiel Dini u. a. (2005); Liebwald (2007); Schweighofer und Geist (2007)). Seit 2006 befasst sich im Rahmen der TREC-Konferenz ein eigener Track mit IR auf der Basis (amerikanischer) Rechtsdokumente (Tomlinson u. a. (2006, 2007)).

Neben IR-Prototypen wurden in einigen Projekten Verfahren zum automatischen Zusammenfassen rechtssprachlicher Dokumente entwickelt. Das SALOMON-System (Moens u. a. (1997)) ermittelt (für Dokumente in einem niederländischen Rechtsprechungskorpus) gestützt auf *cue phrases* zunächst eine grobe Textstruktur, um relevante von irrelevanten Textsegmenten zu unterscheiden. Innerhalb der relevanten Segmente werden dann mit Clustering-Techniken repräsentative Sätze für ein tabellarisches Abstract des jeweiligen Falls ausgewählt. Auch das LetSum-System (Farzindar und Lapalme (2004b,a))

generiert tabellarische Abstracts, nutzt für die Auswahl relevanter Inhalte jedoch eine *thematische* Klassifikation von Textpasssagen. Diese wird regelbasiert auf der Grundlage sprachlicher, struktureller und Layout-bezogener Merkmale ermittelt. Im SUM-Projekt (Grover u. a. (2003); Hachey und Grover (2005, 2004)) wird mit trainierten Klassifizieren experimentiert, die die argumentative Diskursfunktion einzelner Sätze bestimmen sollen.

In den letzten Jahren schließlich befassen sich vermehrt Untersuchungen mit der gezielten Identifikation bestimmter Informationstypen in Rechtstexten. Neben Ansätzen zum automatischen Erwerb ontologischen Wissens (vgl. Lame (2002, 2003); Lame und Desprès (2005), Saias und Quaresma (2003b,a) und Lenci u. a. (2007)) wird dabei auch versucht, gezielt auf bestimmte Typen spezifisch juristischer Information, wie Normen (van Engers u. a. (2004); Biagioli u. a. (2005)) und Argumente (Moens u. a. (2007)), zuzugreifen.

Die genannten Ansätze verwenden dabei zumeist flache linguistische Analyseverfahren. Nur das in Saias und Quaresma (2003b) beschriebene System stützt sich auf semantische Textrepräsentationen, ist jedoch bisher nur partiell implementiert und nicht evaluiert. Einen weiteren – zum Teil evaluierten, jedoch bisher nicht im Kontext eines Gesamtsystems genutzten – Ansatz zur Konstruktion semantischer Repräsentationen für Rechtstexte präsentiert McCarty (2007).

## 3.4 Besonderheiten der Rechtsdomäne

Bei der Entwicklung moderner Zugriffsverfahren für Rechtsinformation wurden bisher weniger greifbare Erfolge erzielt als in manchen anderen Bereichen. Ein historischer Grund hierfür dürfte darin liegen, dass in der Forschung zunächst der Bereich der Expertensystemtechnologie fokussiert wurde, in dem in verschiedenen Projekten vor allem negative Ergebnisse erzielt wurden. Auch die Funktion und Struktur von Rechtstexten weisen jedoch eine Anzahl von Besonderheiten auf, die die Übertragung von Verfahren und Ergebnissen aus anderen Bereichen auf den textbasierten Informationszugriff in der Rechtsdomäne erschweren (vgl. auch Turtle (1995)):

- In sprachlicher Hinsicht unterscheiden sich Rechtstexte durch ihre Komplexität von Texten in den meisten anderen Domänen, in denen Systeme zum automatisierten Informationszugriff genutzt werden. Dies gilt sowohl auf Text- als auch auf Satzebene.

  Verglichen etwa mit Zeitungstext, technischer Dokumentation oder medizinischen Abstracts weisen sowohl Gesetzes- als auch Urteilstexte nicht nur eine größere Länge, sondern vor allem eine komplexere, stärker

konventionalisierte und in hohem Maße bedeutungstragende Dokumentstruktur auf. Auch hinsichtlich der Satzkomplexität und der syntaktischen Komplexität und Variabilität der verwendeten Termini (die etwa eingebettete Relativsätze enthalten und in verbalen und nominalen Varianten auftreten können) stellen Rechtstexte einen Sonderfall dar (vgl. hierzu 4.3.2).

- Die festgelegte Funktion vieler Rechtsdokumente innerhalb gesellschaftlicher Institutionen und Verfahren führt zudem dazu, dass bei der Handhabung textueller Information eine große Vielzahl von Metadaten und juristischem Hintergrundwissen (zumindest potentiell) relevant sind. So ist beispielsweise zu einer abgerufenen Textpassage in beinahe jeder juristischen Anwendung der Zugriff auf Information zur Quelle und deren Status im Bezug auf andere Quellen unverzichtbar. Eine Textpassage aus einer Entscheidung eines Amtsgerichts ist als Fundstelle anders zu bewerten als eine Fundstelle in einem Urteil des BGH.

- Viele Systeme für den textbasierten Informationszugriff bauen auf Redundanzen in der Textgrundlage. Wir haben dies am Beispiel von Question Answering-Systemen erläutert. In der Rechtsdomäne kann jedoch nicht in gleicher Weise auf solche Redundanzen gesetzt werden wie zum Beispiel in Zeitungs- und Nachrichtentexten. Im Extremfall ist weitreichende und normativ bindende Information nur in einem einzigen Dokument enthalten, auf das dann in anderen Dokumenten dann nur noch verwiesen wird (ein gewisses Maß an Redundanz ergibt sich jedoch oft auch hier dadurch, dass Text in Form von Zitaten wiederholt wird).

- Die Rechtssprache ist in einem Kernbereich durch ein hohes Maß an Exaktheit und zudem Verbindlichkeit der Formulierung geprägt: Nicht nur sind alltagssprachlich äquivalente Formulierungen oft juristisch nicht äquivalent zu verwenden. In vielen Fällen ist die *genaue* Einhaltung eines Formulierungsmusters pragmatisch notwendig, um einem Text Geltung zu verleihen. Gleichzeitig zeichnet sich die Rechtssprache allerdings durch die Vermischung rechtsfachsprachlicher Elemente mit Elementen aus der Alltagssprache und den Fachsprachen der jeweiligen Falldomänen aus. Eine Grenze zwischen juristischer Fachsprache und anderen Sprachebenen lässt sich dabei nur schwer (und nicht nach innersprachlichen Gesichtspunkten) ziehen. Somit kommt es für die Verarbeitung juristischer Information in hohem Maße auf eine besonders exakte Analyse der sprachlichen Form an. Sprachtechnologische Ressourcen und Werkzeuge aus anderen Bereichen müssen hierfür jedoch erst angepasst werden.

Diese Besonderheiten legen den Schluss nahe, dass für den Informationszugriff in juristischen Texten sprachlichen Gesichtspunkten, Fragen der sprachtechnologischen Verarbeitung sowie fachspezifischen Gesichtspunkten hohe Aufmerksamkeit zukommen muss. Für den in dieser Arbeit vorgestellten Ansatz zur Extraktion von Definitionen in Urteilstexten können wir uns zum einen auf die im vorigen Kapitel vorgestellte korpusbasierten Analyse des Phänomens juristische Definition in sprachlicher und funktionaler Hinsicht stützen. Zum anderen nutzen wir eine – verglichen mit den meisten in diesem Kapitel diskutierten Systemen – umfangreiche, fachsprachlichen Besonderheiten angepasste linguistische Vorverarbeitung. Auf diese werden wir im folgenden Kapitel näher eingehen.

# Kapitel 4

# Verfahrensschritte, Vorverarbeitung und Ressourcen

In diesem und den folgenden Kapiteln wenden wir uns dem sprachtechnologischen Hauptthema dieser Arbeit zu. Wir befassen uns mit der Entwicklung eines Verfahrens zur automatischen Identifikation und Verarbeitung von Definitionen in Urteilsbegründungen. Der Schwerpunkt liegt dabei zuerst auf dieser Aufgabe als eigenständiger Zielsetzung. Auf die Frage möglicher Anwendungsszenarios kommen wir dann in Kap. 7 zu sprechen.

Wir analysieren zunächst, aus welchen Einzelschritten ein solches Verfahren aufgebaut werden kann und wie diese zusammenwirken. Diese Analyse wird uns in der weiteren Darstellung als Bezugsrahmen dienen. In diesem Kapitel befassen wir uns dann erst näher mit den von uns gewählten Lösungen zur Realisierung der benötigten computerlinguistischen Vorverarbeitungsaufgaben. Im Anschluss charakterisieren wir kurz die Datengrundlage, auf die wir bei der Umsetzung und Erprobung unseres Entwurfs Zugriff hatten. Es handelt sich um einen (gegenüber dem bisher erwähnten Korpus erheblich erweiterten) Bestand von insgesamt mehr als 33 000 Entscheidungstexten. Auch diese Texte hat uns die Firma *juris* freundlicherweise verfügbar gemacht. Abschließend gehen wir in einem Exkurs auf verschiedene interessante Beobachtungen zu linguistischen Eigenheiten der Rechtssprache ein. Unseres Wissens wurden solche Phänomene bisher kaum in größerem Umfang datenbasiert untersucht. Ein Korpus der genannten Größenordnung ermöglicht auf das Thema eine neue, empirische Perspektive. Besondere Beachtung werden wir denjenigen Beobachtungen zukommen lassen, die zu speziellen Anforderungen an die sprachtechnologische Verarbeitung führen.

Im nächsten Kapitel erläutern und evaluieren wir dann die Verfahrensschritte zur Definitionsextraktion und -segmentierung, also die eigentlichen Kernbestandteile unseres Definitionsextraktionsverfahrens.

## 4.1 Gesamtaufbau unseres Definitionsextraktionsverfahrens

Abb. 4.1 gibt einen Überblick über die Einzelschritte unseres Definitionsextraktionsverfahrens und ihre Interaktion. Die mit römischen Zahlen nummerierten Blöcke im Diagramm stehen für die Hauptprozessschritte. Verwendete und erzeugte Datenbestände sind mit Großbuchstaben gekennzeichnet. Die Pfeile repräsentieren den Informationsfluss zwischen den Schritten. Farbig unterlegt sind jeweils einige Angaben zu den von uns für die konkrete Umsetzung der einzelnen Schritte gewählten Technologien und Formaten.

Wie in Kap. 3 näher erläutert liegen Verfahren zum automatischen textbasierten Informationszugriff generell mehrstufige Prozesse zu Grunde. In der Regel ist dem tatsächlichen Informationszugriff eine Phase der Aufbereitung der Textgrundlage vorgelagert, auf deren Ergebnissen dann eine optimierte Suche möglich ist. An die eigentliche Suche schließt sich unter Umständen eine Optimierungsphase an, deren Ergebnisse dann schließlich dem Benutzer verfügbar gemacht werden. Die durch gepunktete Linien angedeuteten Boxen in Abb. 4.1 markieren diese Prozessphasen, mit denen wir uns jeweils die in diesem (Vorverarbeitung), dem nächsten (Suche und Segmentierung) bzw. dem übernächsten Kapitel (Qualitätssicherung und -verbesserung) befassen. Auf einige explorative Experimente zu den in Block V genannten Anwendungsbeispielen gehen wir in Kap. 7 ein.

In vielen Fällen wird ein großer Teil des dargestellten Ablaufs in offline arbeitende Komponenten vorverlagert. Dadurch können verarbeitungsaufwendige Methoden eingesetzt und dennoch Benutzeranfragen effizient behandelt werden. In unserer Implementierung des in Abb. 4.1 dargestellten Verfahrens (die wir im Folgenden kurz als *CORTE-System*[1] bezeichnen werden) haben wir alle Verarbeitungsphasen (also die Blöcke I bis IV im Diagramm) als offline arbeitende Komponenten umgesetzt, um so linguistische Verfahren auch dann verwenden zu können, wenn ihre Performanz für den Echtzeiteinsatz nicht ausreicht.

Für die Implementierung unseres Definitionsextraktionssystems haben wir durchgängig die Programmiersprache Perl und Module aus dem CPAN-Archiv[2] genutzt bzw. auf generell oder zu wissenschaftlichen Zwecken frei verfügbare Lösungen zurückgegriffen. Die einzige Ausnahme stellt das im PreDS-Parser (siehe 4.2.3) integrierte kommerzielle morphologische Analyse-

[1] Die in dieser Arbeit beschriebenen Forschungen wurden zu einem großen Teil im Rahmen des von der Deutschen Forschungsgemeinschaft geförderten Projekts CORTE (**Co**mputerlinguistische Methoden für die **R**echs**te**rminologie) durchgeführt.

[2] Eine umfangreiche Bibliothek von der *community* beigetragener Perl-Module für die verschiedensten Aufgabenstellungen, `http://www.cpan.org`.

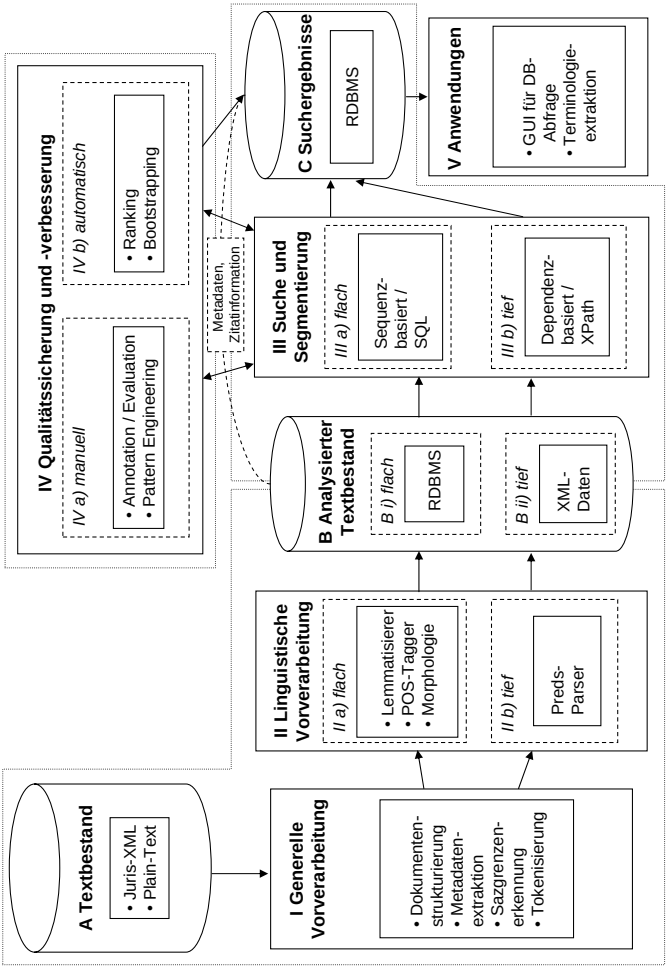



Abbildung 4.1: Einzelschritte unseres Definitionsextraktionsverfahrens

werkzeug Gertwol dar. Der Datenaustausch zwischen den einzelnen Komponenten erfolgt in Form von XML-Strukturen.

## 4.2 Vorverarbeitung

Die in diesem Kapitel diskutierte Vorverarbeitung zerfällt in die generelle Aufbereitung der Eingangsdokumente (I) und ihre anschließende computerlinguistische Aufbereitung (II). Hier teilt sich der Datenfluss in einen tiefen und eine flachen Pfad, auf denen zunächst unterschiedliche sprachtechnologische Werkzeuge für die Korpusaufbereitung (II a bzw. II b) und anschließend auch verschiedene Verfahren für die eigentliche Definitionsextraktion (III a bzw. III b) zum Einsatz kommen.

Die Definitionsextraktion über den flachen Pfad stützt sich auf lemmatisierten und *Part-of-Speech*-getaggten Text. Sie nutzt somit nur sehr elementare, jedoch robust und mit sehr guter Abdeckung erzeugbare linguistische Information. Auf dem tiefen Pfad werden als Extraktionsgrundlage Dependenzstrukturen für den Eingabetext erzeugt, mithin stark strukturierte und normalisierte linguistische Analysen. Diese können zwar nur mit größerem Aufwand, weniger zuverlässig und mit geringerer Abdeckung ermittelt werden. Die in den Strukturen kodierte Information erlaubt jedoch unter Umständen sehr hochwertige Suchergebnisse.

Eine sinnvolle Kombination der beiden Pfade besteht somit in der Nutzung des flachen Pfads als Ergänzung und Fallbacklösung zum tiefen Pfad. Der direkte Vergleich beider Verarbeitungswege erlaubt zudem interessante Einsichten über den Beitrag linguistischer Information verschiedener Tiefe auf die Ergebnisqualität bei einer einheitlichen Informationszugriffs-Aufgabenstellung. Mit diesem Thema werden wir uns in Kap. 5 befassen.

### 4.2.1 Generelle Vorverarbeitung

Viele der in der Literatur beschriebenen experimentellen Systeme für den automatischen Informationszugriff sind für die Evaluation im Rahmen etablierter Wettbewerbe (*shared tasks*) konzipiert. Dabei ist das Format der Eingabedaten in der Regel festgelegt. Zudem handelt es sich bei den zu verarbeitenden Daten meist um relativ kurze und intern nicht tiefer strukturierte Dokumente wie z.B. Zeitungsartikel oder Newswire-Meldungen. Auf eine Vorverarbeitung zur Textnormalisierung und zur Erkennung von Dokumenteigenschaften kann deshalb in diesen Fällen oft weitgehend verzichtet werden.

### (a) Dokumentstruktur von Entscheidungstexten

Bei Gerichtsurteilen ist eine solche Vorverarbeitung dagegen unverzichtbar. Entscheidungstexte sind normalerweise längere Dokumente, und sie weisen eine ausgeprägte und inhaltsrelevante interne Strukturierung auf. In Tabelle 4.2 ist die Grundstruktur eines Entscheidungstexts dargestellt. Sie ist in wesentlichen Punkten durch die Prozessordnungen normiert (z.B. §313 ZPO, §117 VwGO[3]).

---

I. Rubrum

II. Tenor

III. Tatbestand

IV. Entscheidungsgründe

A. Zulässigkeit der Klage

B. Begründetheit der Klage

a) Anspruchsgrundlage

b) Tatbestandsvoraussetzungen der Anspruchsgrundlage
(bzw. der Ermächtigungsnorm im öffentlichen Recht)
(Im Verwaltungsrecht: formelle und materielle Voraussetzungen!)
(1) 1. Tatbestandsmerkmal
(a) Definition
(b) Subsumtion
(2) 2. Tatbestandsmerkmal
...
(3) 3. Tatbestandsmerkmal
...
(n) Letztes Tatbestandsmerkmal

c) Rechtsfolge (Anspruchsinhalt bzw. Reichweite der Ermächtigung, Ermessen)

d) Nebenforderungen (z. B. Zinsen, nicht im öffentlichen Recht)

C. Nebenentscheidungen

1. Kosten

2. Vorläufige Vollstreckbarkeit

---

Abbildung 4.2: Grundstruktur eines Entscheidungstextes

Im Rubrum ist (bis auf die einleitende Formel *Im Namen des Volkes*) kein echter Text enthalten, sondern dokumentarische Informationen angeführt (die Bezeichnung der Prozessbeteiligten, ihrer gesetzlichen Vertreter und der Bevollmächtigten). Bei der Publikation wird das Rubrum in der Regel nicht wörtlich wiedergegeben. Die relevante Information wird vielmehr nach einem durch

[3] Für den Strafprozess sind §§267 und 275 StPO einschlägig, allerdings enthält das von uns genutzte Korpus keine strafrechtlichen Entscheidungen, vgl. 4.3.1.

das jeweilige Publikationsorgan vorgegebenen Schema erfasst. Die restlichen Dokumentbestandteile enthalten dagegen voll ausformulierten Text.

Der Tenor wird mit einem Satz wie *In dem Rechtsstreit... hat... für Recht erkannt* eingeleitet. Er benennt in stark formelhafter Sprache die Rechtsfolge, die das Gericht anordnet (sog. Urteilsformel). Der verbleibende Urteilstext zerfällt in die Beschreibung des Tatbestands und die Erläuterung der Entscheidungsgründe, mit den in Tabelle 4.2 bezeichneten Aufgaben. Diese Teile enthalten frei (wenn auch nach Vorbildern und fachlichen Konventionen) formulierten Text und können von sehr unterschiedlicher Ausführlichkeit sein. Die Textstruktur auf oberster Ebene (Rubrum, Tenor, Tatbestand, Entscheidungsgründe) ist vollständig verbindlich und spiegelt sich praktisch immer in der Dokumentstruktur wider. Dagegen stellt die in Tabelle 4.2 wiedergegebene Einteilung der Entscheidungsgründe zwar eine Anforderung an den Aufbau einer Urteilsbegründung "lege artis" dar. Sie wird aber faktisch nicht immer eingehalten. Sie kann in unterschiedlicher Ausprägung realisiert werden und ist bei weitem nicht in allen Fällen aus der Dokumentstruktur abzulesen.

Das Entscheidungsdokument wird normalerweise ergänzt durch eine einführende Zusammenfassung der Argumentation in (bei der offiziellen Dokumentation erstellten) Leit- bzw. (vom Herausgeber hinzugefügten) Orientierungssätzen. Der Tatbestand wird in publizierten Urteilen häufig nur zusammenfassend wiedergegeben. Bei einem Teil der im CORTE-Korpus enthaltenen Entscheidungen sind auch die Entscheidungsgründe nur abgekürzt erfasst.

Für die Suche nach Definitionen sind – wie schon in Kap. 2 angesprochen – nur die begründungsbezogenen Dokumentteile relevant. Das Rubrum liefert unter Umständen zusätzliche Informationen, die zur Bewertung von Fundstellen nützlich sein können (z.B. die für die Auswertung von Zitaten wichtigen Angaben zu Gericht und Aktenzeichen).

#### (b) Ermittlung und Repräsentation der Dokumentstruktur

Trotz verschiedener Bemühungen hat sich bisher kein einheitliches Format für die maschinenlesbare Erfassung von Gerichtsurteilen etabliert (vgl. 3.3.3). Die Anbieter von Rechtsprechungsdatenbanken bedienen sich für die Datenverwaltung proprietärer Formate. Das CORTE-System verarbeitet XML-formatierte Eingabedokumente, die nach der Saarbrücker DTD (Gantner und Ebenhoch (2001)) oder dem von *juris* verwendeten XML-Format (in dessen Spezifikation wir freundlicherweise Einblick erhalten haben) strukturiert sind. Alternativ dazu kann die Dokumentstruktur von Urteilen im Plain-Text-Format automatisch ermittelt werden, wenn diese zur Markierung der einzelnen Abschnitte die gängigen Überschriften verwenden.

Anhand der XML-kodierten oder automatisch erkannten Dokumentstruktur identifiziert das CORTE-System dann in den Eingabedokumenten Metadaten aus dem Rubrum sowie Leit- und Orientierungssätze und die Entscheidungsgründe. Die begründungsrelevanten Textteile werden in einem nächsten Schritt in Einzelsätze zerlegt. Die Satzgrenzenerkennung ist in Urteilstexten dadurch erschwert, dass im Vergleich zu vielen anderen Textsorten besonders häufig Punkte auftreten, die keine Satzgrenzen markieren. Dies geht vor allem auf zwei Faktoren zurück. Zum einen bedienen sich die Autoren juristischer Texte einer Reihe ansonsten unüblicher und oft irregulär aufgebauter Abkürzungen, die mit einem Punkt abgeschlossen werden, der dann in der Regel keine Satzgrenze anzeigt. Für eine zuverlässige Ermittlung von Satzgrenzen müssen daher spezifisch juristische Abkürzungen erfasst werden.[4]

Zum anderen werden in juristischen Texten besonders ausgeprägt Nummerierungselemente zur Textstrukturierung und in Verweisen eingesetzt. Diese beschränken sich nicht auf Ziffern, sondern können mehrstellige Buchstabenkombinationen enthalten (römische Ziffern in Groß- und Kleinschreibung sowie alphabetische Nummerierungsschemata), die unter Umständen mit einem Punkt abgeschlossen werden. Dieser markiert dann ebenfalls kein Satzende. Für die weitere Verarbeitung müssen solche satzeinleitenden Nummerierungselemente zudem vom restlichen Text separiert werden, da sie in aller Regel von linguistischen Komponenten nicht korrekt verarbeitet werden können.

Die Satzgrenzenerkennung in unserem System verwendet ein frei verfügbares Programmmodul aus der CPAN-Library, das wir um eine umfangreiche Liste klassifizierter juristischer Abkürzungen, eine Komponente zur Erkennung von Nummerierungselementen sowie um Regeln zur nachträglichen Korrektur fehlerhafter Erkennungsergebnisse ergänzt haben. Als praktisches Hauptproblem bei der Satzgrenzenerkennung haben sich komplexe Normbelege erwiesen.

Abschließend werden die ermittelten Eingabesätze tokenisiert (also in Einzelwörter zerlegt). Die Tokenisierung erfolgt ebenfalls unter Verwendung einer umfassenden Abkürzungsliste sowie heuristischer Regeln, um die korrekte Behandlung verschiedener Sonderfälle (z.B. Nummerierungen und verschiedene Arten der Paragraphenbezeichnung in Rechtsquellenangaben) sicherzustellen.

[4] Eine zusätzliche Schwierigkeit stellt die Tatsache dar, dass bestimmte juristische Abkürzungen (insbesondere Kurzbezeichnungen für Gerichte und Gesetze), in der Regel nicht mit einem Punkt abgeschlossen werden. Diese enthalten oft wortinterne Großbuchstaben und werden daher von den meisten Satzgrenzenerkennern der Wortgestalt nach als Abkürzungen klassifiziert. Ein nachfolgender Punkt gehört jedoch nicht zur Abkürzung, sondern markiert in jedem Fall eine Satzgrenze. Ein in juristischem Text aufgrund der Anzahl von Auftreten sehr relevanter Sonderfall ist die Zeichenfolge *Art*: Es kann jeweils nur aufgrund des Kontextes entschieden werden, ob es sich um das Nomen *Art* oder die Abkürzung des Wortes *Artikel* handelt.

Die aufbereiteten Eingabedokumente werden zum einen in eine Indexdatenbank[5] eingespeist (die somit im Gegensatz zu den verfügbaren Rechtsprechungsdatenbanken eine Adressierung nicht nur auf Dokumentebene, sondern zusätzlich auf Token- und Satzebene ermöglicht). Zum anderen werden sie als XML-Daten an die flache und tiefe linguistische Verarbeitung weitergereicht.

### 4.2.2 Flache Verarbeitung

Die flache linguistische Verarbeitung beeinhaltet eine lexikalische Analyse des Textbestandes. Die Token der Einzelsätze werden lemmatisiert, mit Wortklasseninformation versehen (*Part-of-Speech*-Tagging) und in morphologische Bestandteile zerlegt.

#### (a) Lemmatisierung und *Part-of-Speech*-Tagging

Die Lemmatisierung und das *Part-of-Speech* (POS)-Tagging erfolgen mittels des an der Universität Stuttgart entwickelten Tools TreeTagger (Schmid (1994)), eines sprachunabhängigen probabilistischen POS-Taggers, der für das Deutsche das Stuttgart-Tübinger Tagset nutzt (Schiller u. a. (1999)). Zur Lemmatisierung verwendet TreeTagger ein Vollformenlexikon. Für nicht im Lexikon enthaltene Wortformen wird zwar kein Lemma ausgegeben, es kann jedoch eventuell dennoch (durch Verwendung eines Suffixlexikons oder allein aufgrund der Analyse der Kontextwörter) eine Tagging-Hypothese erzeugt werden.

Sowohl das deutsche Vollformenlexikon als auch das deutsche Sprachmodell wurden auf der Basis von Zeitungstext erzeugt. Für rechtssprachliche Texte kann deshalb nicht ohne weiteres eine ähnliche Performanz erwartet werden wie die in Schmid (1994) berichteten 97,5% Akkuratheit. Eine von uns untersuchte Stichprobe von 1000 analysierten Tokens aus einem Entscheidungstext enthielt 51 falsch getaggte Tokens, bei denen in 38 Fällen zudem kein Lemma identifiziert werden konnte. Bei 40 weiteren Tokens wurde zwar kein Lemma ermittelt, jedoch die korrekte POS-Klassifikation. Sämtliche erkannten Lemmata waren korrekt. Die Akkuratheit in der Stichprobe lag somit bei 94,9% für die POS-Klassifikation, bei 92,2% für die Lemmatisierung und bei 90,9% für die kombinierte Aufgabenstellung. Eine genauere Untersuchung zeigt jedoch, dass weit über die Hälfte der Fehlanalysen in unserer Stichprobe auf spezielle Abkürzungen, verbliebene Nummerierungselemente und Fehler im Eingabetext (z.B. fehlende Leerzeichen) zurückzuführen sind. Betrachtet man nur die

[5] Hier wie bei den im Folgenden erwähnten Datenbanken verwenden wir das frei verfügbare relationale Datenbanksystem MySQL.

Analysen echter Wörter, so ergibt sich eine Akkuratheit von 99,9% für die POS-Klassifikation und 95% für die Lemmatisierung.

Zur Ermittlung von Lemmata für die verbleibenden unanalysierbaren Wortformen nutzen wir in einem nachgelagerten Verarbeitungsschritt das kommerzielle Morphologieanalyse-Werkzeug Gertwol[6] der Firma Lingsoft. Dieses weist zwar eine sehr hohe Abdeckung auf, ermittelt jedoch für Eingaben sämtliche möglichen Analysen (und damit in einem großen Teil der Fälle mehrere mögliche Grundformen). Es ist daher alleine nicht für die Lemmatisierung geeignet. Die durch TreeTagger verfügbar gemachte POS-Information ermöglicht jedoch eine zuverlässige Desambiguierung zwischen alternativen Lemmata. In der von uns betrachteten Stichprobe konnten nach der ergänzenden Analyse durch Gertwol sämtliche echten Wörter korrekt lemmatisiert werden.

#### (b) Morphologische Zerlegung

Im Anschluss an die Lemmatisierung erfolgt auf dem flachen Verarbeitungspfad noch eine auf heuristische Regeln gestützte Zerlegung der Einzelwörter in Wurzelmorpheme. Diese bedient sich der von Gertwol gelieferten Informationen zu Morphemgrenzen sowie einer Anzahl zusätzlicher, in Gertwol nicht implementierter Regeln zur Identifikation von Derivationsmorphemen, um eine über Wortklassengrenzen und Kompositabildung hinweg einheitliche und vergleichbare Wortrepräsentation zu erzeugen (so dass z.B. für die Wörter *immigrieren*, *Immigrant*, *Immigrationsbehörde* und *Non-Immigrant-Visum* der gemeinsame morphologische Bestandteil *immigr* erkannt werden kann). Eine solche Zerlegung würde prinzipiell etwa bei einer definitionsbasierten Dokumentindizierung die Zusammenfassung nominaler und verbaler Repräsentationen desselben Konzepts (z.B. *Immigration* und *immigrieren*) zu einem Indexterm unterstützen. Im Rahmen der hier beschriebenen Arbeiten nutzen wir die ermittelte Information beim Ranking von Treffern der Definitionssuche für eine ähnlichkeitsbasierte Bewertung der Suchergebnisse anhand eines Definitionsprofils[7] (Kap. 6). Die Ergebnisqualität der morphologischen Zerlegung haben wir – zum einen aufgrund der relativ untergeordneten Rolle innerhalb des Gesamtsystems, zum anderen aufgrund des Mangels an vergleichbaren Ansätzen und Referenzressourcen – nicht evaluiert.

[6] **Ger***man* **two-l***evel morphology*

[7] vgl. zur Erläuterung des Begriffs 3.2.3

### 4.2.3 Parsing

Die tiefe linguistische Verarbeitung beeinhaltet eine syntakto-semantische Analyse des Eingabetextes, in die u.a. die Identifikation von domänenspezifischen *Named Entities* (vgl. zur Definition des Begriffs Kap. 3) integriert ist.

#### (a) Anforderungen

Optimale Voraussetzung für einen zielgenauen und qualitativ hochwertigen automatischen Informationszugriff in Textdokumenten wäre eine tiefe und vollständige linguistische Analyse. Die Informationssuche könnte dann gegebenenfalls über vereinfachten Strukturen erfolgen, in die selektiv nur die relevanten Bestandteile eines vollständigen Analyseergebnisses übernommen würden. Gegenwärtig verfügbare Systeme zur tiefen linguistischen Verarbeitung sind jedoch mit teilweise gravierenden Problemen für den praktischen Einsatz in größerem Maßstab behaftet. Dies gilt sowohl für die Performanz als auch für die Robustheit des Analyseprozesses. Performanzprobleme liegen zum Beispiel im Laufzeit- und Speicherbedarf, in mangelnder Abdeckung oder der massiven Generierung von Ambiguitäten (v.a. bei regelbasierten Systemen auf der Basis großer Grammatiken). Robustheitsprobleme betreffen die Fehlertoleranz und Stabilität, also den Umgang mit unerwartetem, fehlerhaftem oder aus Komplexitätsgründen nicht regulär verarbeitbarem Eingabetext.

Bei Rechtsdokumenten resultieren spezielle Robustheitsanforderungen neben der großen Komplexität v.a. aus der Komplexität des Satzbaus und der Verwendung von Fachvokabular aus der Rechtssprache sowie den Domänen der jeweils behandelten Fälle (bei *Agio* in 4.1 etwa handelt es sich um einen wirtschaftssprachlichen Fachbegriff). Hinzu kommen unübliche und teilweise nur im lokalen Kontext gültige Abkürzungen (wie *Astin* für *Antragstellerin*[8] im selben Beispiel) und die gängige Praxis, Angaben zu Rechtsquellen und Belegstellen mit syntaktischen Mitteln in den Text zu integrieren (vgl. 4.2).

(4.1) Zur Behandlung des Agios durch das FA macht die Astin geltend…

(FG 3. Senat, 13. April 1977, AZ III 359/76, juris)

(4.2) Auf das Asylrecht des Art. 16a Abs. 1 GG kann sich gemäß Art. 16a Abs. 2 Satz 1 GG, § 26a Abs. 1 Sätze 1 und 2 AsylVfG nicht berufen, wer…

(Sächsisches Oberverwaltungsgericht 4. Senat, 28. August 2001, AZ A 4 B 4388/99, juris)

[8] Wie oben erwähnt ist die Verwendung von Abkürzungen ohne abschließenden Punkt in Urteilstexten nicht ungewöhnlich.

Schwierigkeiten bereiten außerdem Erfassungsfehler (Eingabefehler, OCR-Fehler, fehlerhafte Realisierung typographischer Hervorhebungen wie Sperrungen und Unterstreichungen), Anonymisierungen durch Abkürzung, Schwärzung oder Auslassung von Eigennamen und schließlich die relativ große Häufigkeit ungrammatischer Sätze (die angesichts der sprachlichen Komplexität nicht überrascht).

Für den praktischen Einsatz zur Verarbeitung großer Mengen unkontrollierter Eingabedokumente ist es unerlässlich, dass ein System auch bei solchen Schwierigkeiten den Analyseprozess fortsetzt und zumindest partielle Ergebnisse für die unproblematischen Bestandteile der Eingabe erzeugt. In den letzten Jahren hat sich eine Vielzahl von Ansätzen etabliert, die darauf abzielen, Praxistauglichkeit durch eine Balance zwischen linguistischer Analysetiefe und -vollständigkeit auf der einen sowie Robustheit und Konsistenz der Analysen auf der anderen Seite zu erzielen.

Dabei werden zum einen auf annotierten Daten trainierte statistische Verfahren genutzt, so z.B. in Charniak (1999) für das Englische und Dubey (2005) für das Deutsche. Mit diesen kann Robustheit dadurch erzielt werden, dass die binäre Entscheidung für oder gegen eine bestimmte Analyse durch eine gradierte Bewertung verschiedener Analysehypothesen ersetzt wird. Zudem lässt sich die Ergebnisqualität statistischer Verfahren bis zu einem gewissen Grad durch die Nutzung weiterer Trainingsdaten erhöhen, deren Gewinnung im Vergleich zum manuellen *Grammar Engineering* weit weniger linguistische Expertise erfordert. Bei Verfügbarkeit entsprechender annotierter Trainingsdaten[9] ist auch die Anpassung an spezifische Domänen möglich.

Zum anderen kommen Kombinationen verschiedener spezialisierter Module zum Einsatz. Hybride Architekturen versuchen, die Probleme tiefer Analysekomponenten durch die ergänzende Nutzung flacher Verfahren zur Vorverarbeitung, als Fallback oder als Informationsquelle bei Entscheidungen im Analyseprozess auszugleichen, vgl. etwa Daum u. a. (2003). Andere Systeme erzeugen aus den Ergebnissen mehrerer flacher Verfahren eine kombinierte Gesamtanalyse (das Grundkonzept solcher Systeme ist beispielsweise in Abney (1996) beschrieben).

Wir verwenden zur tiefen linguistischen Verarbeitung im CORTE-System einen Parser, der den letztgenannten Ansatz verfolgt. Er erzeugt Textrepräsentationen in Form sogenannter *Partially Resolved Dependency Structures* (PreDS). Der PreDS-Parser beruht auf einer kaskadierten Kombination verschiedener relativ einfacher Komponenten. Konzeption und Implementierung des *PreDS*-Parsers sind in Fliedner (2007) und Braun (2003) ausführlich beschrieben und im Hinblick auf den Stand der Technik in der Computerlinguistik diskutiert.

[9] die allerdings für die deutsche Rechtssprache nicht gegeben ist.

Für den Einsatz des PreDS-Parsers im CORTE-System sprach vor allem, dass der im PreDS-Format (das wir im nächsten Abschnitt genauer beschreiben) realisierte Grad der Abstraktion von der sprachlichen Oberflächenform für die anvisierte Aufgabenstellung besonders geeignet ist.

Zudem belegten Erfahrungen aus anderen Projekten die technische Eignung des PreDS-Parsers. Er wurde bereits im COLLATE-Projekt zur Analyse von Zeitungstext genutzt und wird außerdem in einem linguistisch informierten QA-System verwendet (letzteres ist in Fliedner (2007) beschrieben). Zu Beginn der Arbeiten am CORTE-System war kein vergleichbares Parsingsystem für das Deutsche verfügbar, das bereits ähnlich erfolgreich im Rahmen des textbasierten Informationszugriffs eingesetzt worden war.

Auch die hohen Robustheitsanforderungen, die aus der sprachlichen Komplexität und den sonstigen Besonderheiten unserer Eingabedaten resultieren, erfüllte der PreDS-Parser von Anfang an in vergleichsweise hohem Maße. Ein zu Beginn unserer Arbeiten durchgeführtes Experiment mit zwei wichtigen damals vefügbaren weiteren Parsern für das Deutsche, dem statistischen Sleepy-Parser (Dubey (2005)) und dem LFG-basierten XLE-System (Dipper (2003), Maxwell und Kaplan (1991)), hat ergeben, dass mittels des PreDS-Parsers ohne weitere Anpassungen ein erheblich höherer Anteil der Sätze in unserem Pilotstudien-Korpus (vgl. Kap. 2) analysiert werden konnte als mit den beiden anderen Systemen (3105 der insgesamt 3509 Sätze, verglichen mit 2453 für den Sleepy-Parser und 1025 für das XLE-System, siehe Walter (2009)). Während für den Sleepy-Parser vor allem Timeouts und ein aus Ressourcengründen festgelegtes Satzlängen-Maximum zu Problemen führten, konnte das XLE-System eine Vielzahl spezieller Ausdrücke im Text (etwa die bereits angesprochenen Angaben zu Belegstellen bei Zitaten) nicht verarbeiten.

Zudem bot der PreDS-Parser uns die Möglichkeit, selber problemlos Anpassungen an der Implementierung vorzunehmen, um das System so für die Verarbeitung juristischer Texte zu optimieren (vgl. Abschnitt d). Möglicherweise hätten ähnliche Anpassungen auch bei anderen Systemen zu einer deutlichen Verbesserung der Leistung führen können. Sie wären aber in jedem Fall mit erheblich höherem Implementierungsaufwand verbunden gewesen. Beim heutigen Stand der Entwicklung könnte eine in einem hybriden Framework wie Heart of Gold (Schäfer (2007)) integrierte Kombination aus spezialisierten Vorverarbeitungskomponenten und einem großen generischen Grammatiksystem (etwa der in Crysmann u. a. (2002) und Müller und Kasper (2000) beschriebenen HPSG für das Deutsche) eine aus technischer Sicht gangbare Alternative zur Verwendung des PreDS-Parsers darstellen.

Wir beschreiben im Folgenden zunächst genauer das PreDS-Format und dann den Aufbau und die Komponenten der im PreDS-Parser umgesetzten kaskadierten Systemarchitektur. Für eine umfassendere Beschreibung verweisen

wir wiederum auf Fliedner (2007). Schließlich gehen wir kurz auf die von uns vorgenommenen Anpassungen zur Analyse von juristischem Text und auf einige Daten zur Qualität der Analyseergebnisse ein.

#### (b) Das PreDS-Format

Bei dem an der Universität des Saarlandes entwickelten Textrepräsentationsformat *Partially Resolved Dependency Structures* (PreDS) handelt es sich um eine stark normalisierte Form semantisch orientierter Dependenzstrukturen.

In Abb. 4.3 sind für den Satz in (4.3) jeweils eine typische konstituentenstrukturelle (A) und dependenzgrammatische Analyse (B) sowie die PreDS-Repräsentation (C) dargestellt. Die Darstellung der Analysen A und B ist schematisch, denn verschiedene Ausprägungen beider grammatiktheoretischer Ansätze weichen teils erheblich voneinander ab.

(4.3) Eine unrichtige Umsetzung kann nur angenommen werden, wenn der inhaltliche Spielraum überschritten ist.

(modifiziert aus: Oberverwaltungsgericht des Saarlandes 8. Senat, 10. März 1995, AZ 8 N 5/95, juris)

In Konstituentenstrukturanalysen werden die satzinternen syntaktischen Relationen als hierarchische Beziehungen zwischen abstrakten Kategorien aufgefasst, auf die die Wörter des Satzes projiziert werden. Dependenzstrukturen beschreiben die grammatische Struktur von Sätzen dagegen durch gerichtete hierarchische Relationen (Dependenzen) zwischen den Wörtern (oder größeren Einheiten wie den Verbalkomplexen in Abb. 4.3) selber. Im Gegensatz zu Konstituentenstrukturgrammatiken erzeugen Dependenzgrammatiken also Baumstrukturen, deren Blätter *und* innere Knoten direkt zu Wörtern des Satzes korrespondieren. Die Baumstruktur ist in dependenzgrammatischen Analysen unabhängig von der Oberflächenstruktur des jeweiligen Satzes und liegt näher an dessen semantischer Struktur. Im Gegensatz zu den meisten konstituentenstrukturellen Analysen wird die Oberflächenreihenfolge der Wörter nicht in der Baumstruktur wiedergegeben.

In PreDS werden im Unterschied zu vielen anderen Typen dependenzgrammatischer Analysen nur die Lemmata der Inhaltswörter eines Eingabesatzes als Knoten im Dependenzbaum repräsentiert. Zudem beschränken sich die Kantenlabels auf eine relativ kleine Menge von Bezeichnern elementarer tiefengrammatischer Relationen. Die von Funktionswörtern beigetragene Information wird in Merkmalsstrukturen kodiert, die mit den Knoten bzw. Kanten verknüpft sind. Das verwendete Relationsinventar ist in Tabelle 4.1 angegeben. Zu den Modifikator-Relationen existieren jeweils Varianten mit dem Suffix *Def*

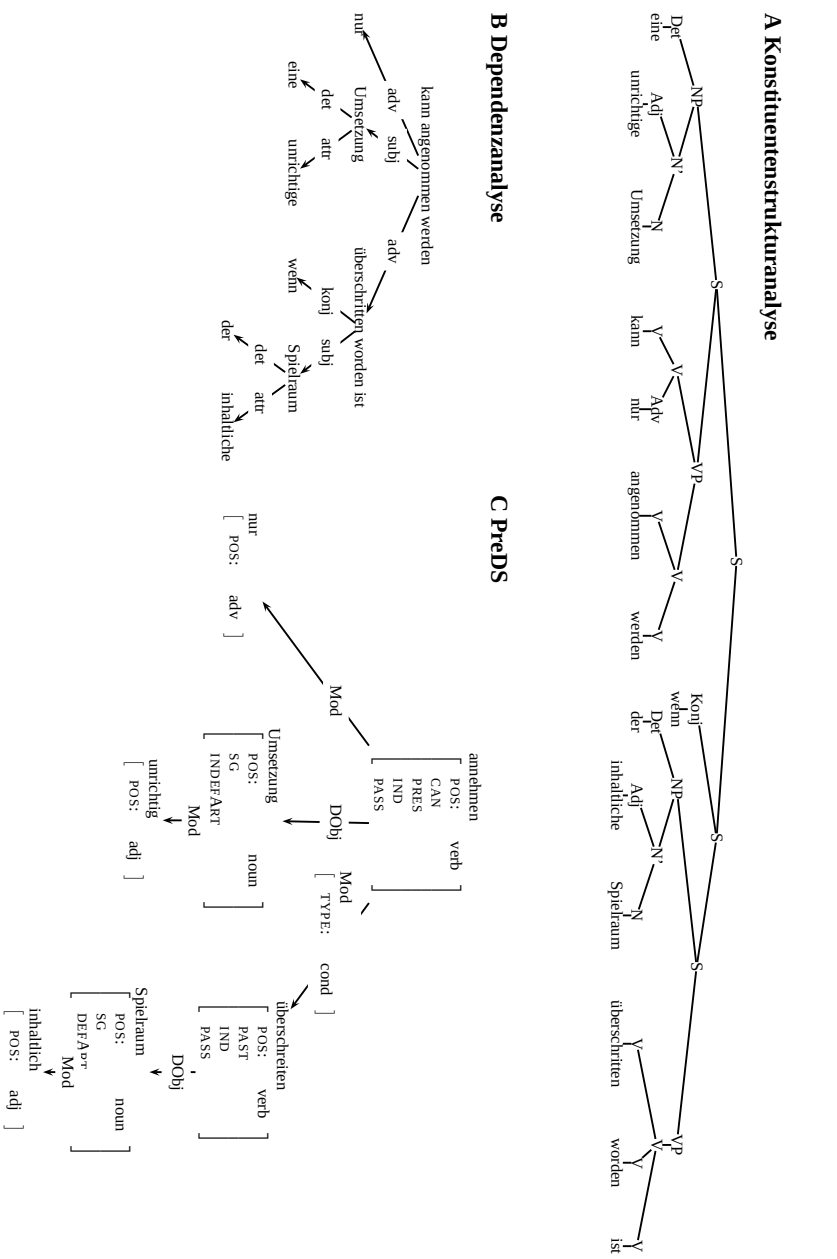



Abbildung 4.3: Vergleich einer Konstituentenstruktur-, Dependenz- und PreDS-Analyse für Beispiel (4.3)

zur Kennzeichnung von unterspezifizierter Default-Anknüpfung. Wir gehen in Abschnitt (c) noch auf die Verwendung diese Mechanismus im PreDS-Parser ein.

| **Relation** | **Beschreibung** | **Beispiel** |
|---|---|---|
| DSub | Tiefensubjekt | *Der Kläger* spricht. |
| DNom | Prädikatsnomen | Er ist *Kläger*. |
| DObj | Tiefenobjekt | Er ruft *den Kläger*. / *Der Kläger* wird gerufen. |
| DInd | Indirektes Tiefenobjekt | Er schreibt *dem Kläger*. |
| DGen | Tiefen-Genitivobjekt | Es bedarf *eines Klägers*. |
| DSub2 | Nominalkomplement zum Tiefensubjekt | Er gilt *als Kläger*. |
| DObj2 | Nominalkomplement zum Tiefenobjekt | Sie betrachten ihn *als Kläger*. |
| ClArg | Satzwertiges Komplement | Es ist notwendig, *dass der Kläger anwesend ist*. |
| Mod | Generelle Modifikatoren (z.B. Attribute, Adverbien, Relativsätze) | Der *erste* Kläger spricht. |
| PPMod | Präpositionalgruppen | Die Reaktion *auf seine Klage* ist offen. |
| NPMod | Modifizierende Nominalgruppen | Er *als Kläger* muss teilnehmen. |
| GenMod | Genitivattribute | Die Teilnahme *des Klägers* ist notwendig. |
| NPArg | Funktional unterspezifizierte Nominalgruppen | |

Tabelle 4.1: PreDS-Relationen

Das PreDS-Format erlaubt somit eine (auch im Vergleich zu anderen dependenzgrammatischen Formaten) sehr weitgehende Normalisierung unterschiedlicher Oberflächenrealisierungen gleicher oder verwandter Inhalte. Im Folgenden sind die wichtigsten Arten von Formulierungsvarianten beschrieben, für die im PreDS-Format strukturell einheitliche Repräsentationen verwendet werden:

**Diathesen:** Die Unterscheidung korrespondierender Aktiv- und Passiv-Konstruktionen erfolgt durch ein Merkmal [pass] am Knoten des be-

treffenden Prädikats. Dieser enthält das Lemma des Hauptverbs, das Hilfsverb wird nicht repräsentiert. Über die Valenzalternation wird abstrahiert, korrespondierende Argumente in Aktiv- und Passiv-Varianten sind durch identische Dependenzrelationen mit dem übergeordneten Prädikat verknüpft. Bei Reflexivkonstruktionen wird am Prädikatknoten das Merkmal [reflex] gesetzt.

**Tempus:** Auch das Tempus wird durch entsprechende Merkmale am Prädikatknoten markiert (z.B. [praes] oder [fut]). Bei zusammengesetzten Tempusformen enthält der Prädikatknoten wiederum das Lemma des Hauptverbs. Das Hilfsverb (bzw. die Hilfsverben) taucht / tauchen in der PreDS nicht auf.

**VP-Negation, Modus und Modalisierung:** Negation, Modus und modale Hilfsverben werden ebenfalls als Merkmale des Prädikats erfasst. Gegebenenfalls werden für ein Hilfsverb mehrere Merkmale gesetzt (z.B. [konj] und [can] für *könnte*). Die in Rechtstexten besonders häufige modale Passivumschreibung mit *ist zu* wird als Passiv analysiert und zusätzlich durch das Merkmal [modSein] markiert.

**Adjektivprädikationen:** Sätze mit prädikativ gebrauchten Adjektiven werden strukturell analog zu Sätzen mit rein verbalem Prädikat analysiert: Das Adjektiv wird als Lemma des Prädikatknotens dargestellt, die Kopula trägt Modusmerkmale bei und wird nicht repräsentiert.

**Definitheit / Artikel:** Knoten mit nominalem Lemma tragen (entsprechend dem verwendeten Artikel, der selbst nicht repräsentiert wird) ein Merkmal [defArt] oder [indefArt]. Determinativ verwendete adjektivische Demonstrativpronomen (z.B. *diese Verordnung* oder *alle Ansprüche*) werden als attributive Adjektive repräsentiert und tragen ein Merkmal [detArt] zum Knoten des Bezugsworts bei.

**Modifikation:** Attributive Adjektive, adverbiale Bestimmungen, attributive Relativsätze und Adverbialsätze werden einheitlich durch die Dependenzrelation *Mod* (bzw. ihre durch ein Präfix gekennzeichneten Subtypen) mit ihrem jeweiligen Bezugswort verknüpft. Diese wird durch entsprechende zusätzliche Merkmale näher bestimmt. Bei modifizierenden Präpositionalphrasen wird in einem zusätzlichen Merkmal die einleitende Präposition angegeben.

**Abtrennbare Wortbestandteile:** Abtrennungen von Verbpräfixen sowie Kompositionserstgliedern (in Aufzählungen) werden in der PreDS rückgängig gemacht.

**Freie Relativsätze:** Freie Relativsätze (d.h. mit *wer*, *was*, *wen* usw. eingeleitete Relativsätze, die als Subjekt oder Objekt des übergeordneten Satzes fungieren) werden strukturell wie ein entsprechender nominaler Satzteil behandelt. Sie werden über die entsprechende PreDS-Relation angebunden.

Neben einer vollständigen Normalisierung des Verbalkomplexes leistet das PreDS-Format also auch eine weitgehende Vereinheitlichung nominaler Konstruktionen. Für den automatisierten Informationszugriff stellen die angesprochenen Normalisierungen eine wichtige Vereinfachung dar. Sie abstrahieren nämlich über die Variabilität des sprachlichen Ausdrucks in soweit, als diese für den eigentlichen Kerninhalt irrelevant ist. Semantisch äquivalente oder ähnliche Informationseinheiten können deshalb in PreDS einheitlich, zum Beispiel durch gemeinsame Suchmuster, identifiziert und mit den gleichen Mechanismen verarbeitet werden.

#### (c) Der PreDS-Parser

Wir gehen nun kurz auf den Aufbau des Parsers sowie einige Modfikationen und Ergänzungen ein, die erforderlich waren, um die Analysequalität in den für unsere Anwendung relevanten Bereichen zu erhöhen.

In Abb. 4.4 ist die Abfolge der einzelnen Systemkomponenten dargestellt. Jede der Komponenten korrespondiert jeweils zu einem Typ linguistischer Strukturen und hat Zugriff auf die Analyseergebnisse der vorgelagerten Komponenten (also auf linguistische Strukturen untergeordneter Ebenen). Die ermittelten Analysen werden von Komponente zu Komponente XML-kodiert weitergegeben. Jeder Verarbeitungsschritt erweitert die Eingabestruktur inkrementell um weitere Analyseergebnisse.

**Tokenisierung.** Die Tokenisierung stützt sich auf das das bereits in 4.2.1 angesprochene Modul, berücksichtigt also ebenfalls eine Vielzahl von Abkürzungen und Ausnahmefällen.

**Morphologische Analyse.** Die morphologische Analyse bestimmt sämtliche morphologischen Lesarten der Token des Eingabesatzes. Sie erfolgt durch das (bereits in 4.2.2 erwähnte) kommerzielle Werkzeug *Gertwol*. Es basiert auf dem Ansatz der *Zwei Ebenen-Morphologie*, beeinhaltet ein Stammlexikon mit etwa 350 000 Einträgen und deckt die Flexionsparadigmata, Derivationsschemata und Regeln zur Kompositabildung des Deutschen mit großer Vollständigkeit ab. Sämtliche morphologischen Lesarten werden mit ihren morphosyntaktischen Merkmalen den weiteren Verarbeitungsschritten verfügbar gemacht.

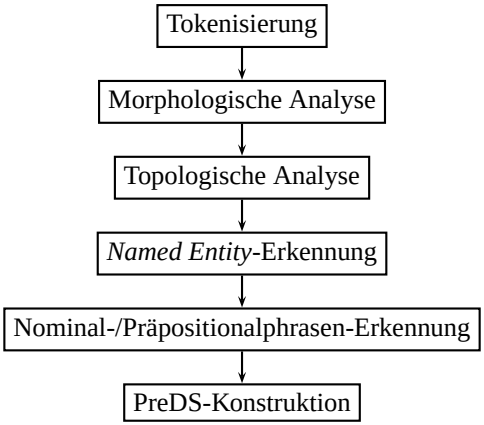



Abbildung 4.4: Systemkomponenten des PreDS-Parsers, nach (Fliedner, 2007, S. 249)

**Topologische Analyse.** Im nächsten Schritt wird unter Verwendung des von Christian Braun an der Universität des Saarlandes entwickelten topologischen Parsers (Braun (1999)) die Grobstruktur der Haupt- und Nebensätze in der Eingabe ermittelt. Die hierfür genutzte kontextfreie Grammatik umfasst etwa 350 Regeln und deckt den Großteil der im Deutschen möglichen Satztypen ab.

Der topologische Beschreibungsansatz geht auf das vorrangig von Drach (1937) eingeführte *Stellungsfeldermodell* zurück. Nach diesem wird der deutsche Satz durch die sogenannte *linke* und *rechte Satzklammer* in drei Stellungsfelder (*Vor-*, *Mittel-* und *Nachfeld*) unterteilt. Verschiedene Satzbaupläne unterscheiden sich dadurch, ob und wie diese Regionen besetzt sind. Die Satzklammern enthalten normalerweise die Prädikatbestandteile, in Nebensätzen kann die linke Satzklammer die Subjunktion beinhalten. Nach der Position des finiten Verbs unterscheidet man drei Haupt-Satzbaupläne: Verbzweit-Sätze (V2) sind Sätze, in denen das Vorfeld besetzt ist und das finite Verb in der linken Klammer steht. Verbletzt-Sätze (VL) enthalten das finite Verb in der rechten Klammer. In Verberst-Sätzen (V1) ist das Vorfeld unbesetzt und das finite Verb steht in der linken Klammer. In (4.4) bis (4.6) sind Beispiele für diese drei Satzbaupläne angegeben. Die Feldstruktur ist durch eckige Klammern markiert, deren Indizes abkürzend für die jeweilige Satzregion stehen. Alle drei Sätze sind

dem Urteil 3. Zivilsenats des Hanseatischen Oberlandesgerichts Hamburg vom 7. März 2002 (3 U 353/01, juris) entnommen.

(4.4) *V2*: [$_{VF}$ Die Wiederholungsgefahr] [$_{LK}$ wird] [$_{MF}$ auf Grund der Zuwiderhandlung] [$_{RK}$ vermutet].

(4.5) *VL*: . . . , [$_{LK}$ ob] [$_{MF}$ eine entstandene Wiederholungsgefahr] [$_{RK}$ ausgeräumt ist]

(4.6) *V1*: [$_{LK}$ Wird] [$_{MF}$ der. . . erforderliche Hinweis. . . so] [$_{RK}$ gestaltet][$_{NF}$ , dass er sich nur auf eine Patientengruppe . . . bezieht] (, verstößt dies gegen. . . )

Innerhalb der Satzfelder identifiziert der PreDS-Parser dann zunächst *Named Entities* (Eigennamen und vergleichbare Bezeichnungen, deren Aufbau Eigengesetzlichkeiten folgt, die nicht zur allgemeinsprachlichen Grammatik zu zählen sind) und im Anschluss einfache und komplexe Nominal- und Präpositionalphrasen (*Named Entity Erkennung* und *NP / PP-Chunking*).

***Named-Entity*-Erkennung.** Die *Named-Entity*-Erkennung basiert auf Information aus der morphologischen Analyse sowie umfangreichen Speziallexika, in denen einfache und unveränderliche Bestandteile von Bezeichnungen aus verschiedenen Sachbereichen aufgeführt sind (sog. *Gazeteers*). Diese werden mit einer Grammatik auf der Basis regulärer Ausdrücke für Blattknoten kombiniert. Sie beschreibt die Einbettungsmöglichkeiten dieser Fixbestandteile in komplexe Ausdrücke.[10] So dienen zum Beispiel bei der Erkennung von Normverweisen (etwa *Art. 16a Abs. 1 GG* in (4.2)) Normtitel (wie *GG*) als Anker. In deren Umgebung werden detaillierte Stellenangaben nach verschiedenen Gliederungsschemata (z.B. Artikel und Absatz) durch entsprechende reguläre Ausdrücke lizensiert.

**Nominal-/Präpositionalphrasen-Erkennung.** Für die Identifikation von Nominal- und Präpositionalphrasen kommt ein von Gerhard Fliedner entwickelter Phrasen-Chunker (Fliedner (2002, 2001)) auf Grundlage eines erweiterten endlichen Automaten mit einer optimalitätstheoretischen Komponente zur Ambiguitätsauflösung zum Einsatz. Neben einfachen Nominal- und Präpositionalphrasen erkennt diese Komponente auch pränominale Modifikatoren mit eingebetteter Struktur und verarbeitet in Konstituenten integrierte *Named Entities* korrekt.

[10] Die Grammatik nutzt die Ausdrucksmöglichkeiten, die das zur Verarbeitung genutzte CPAN-Modul Parse::RecDescent (`http://search.cpan.org/dist/Parse-RecDescent/`) zur Verfügung stellt.

**Preds-Konstruktion.** Die eigentliche PreDS-Konstruktion erfolgt durch eine Abfolge von Operationen, die eine Kaskade heuristischer Regeln umsetzt. Dabei werden für jede durch den topologischen Parser identifizierte klausale Struktur zunächst aus dem Material in den Verbklammern Lemma und Attribute des Prädikatknotens ermittelt. Dann werden Nominalphrasen und Präpositionalphrasen in den Feldern zu möglichst großen Einheiten zusammengefasst. Dabei werden auch Dependenzkanten auf der Grundlage postnominaler Modifikation erzeugt. Die erzeugten Einheiten füllen dann die Argumentstellen des Prädikats. Abschließend werden Nebensätze und verbleibendes Material als Modifikatoren an den Prädikatknoten angeknüpft.

Die Verarbeitung erfolgt strikt sequentiell, ein Rücksprung zu Entscheidungspunkten und das Verfolgen von Alternativanalysen ist (bis auf wenige Ausnahmen innerhalb der Grenzen einzelner Module) nicht möglich. In verschiedenen Fällen syntaktischer Ambiguität (z.B. bei Attachement-Ambiguitäten oder unklaren Argumentzuweisungen aufgrund von Kasus-Ambiguitäten) werden Default-Entscheidungen getroffen und als unterspezifiziert markiert. Für unterspezifizierte nominale Komplemente wird die Relation *NPArg* verwendet. Auch zu den Modifikator-Relationen existieren – wie bei Tabelle 4.1 angemerkt – unterspezifizierte Varianten, die jeweils durch das Suffix *Def* (für *Default*) gekennzeichnet sind. Durch die lokal orientierte Verarbeitungsstrategie kann die Komplexität des Analysevorgangs gegenüber einer vollständigen tiefen Analyse stark reduziert werden. Dadurch wird ein Zugewinn an Performanz und Robustheit bei komplexen Eingaben erzielt. Weiterhin können in allen Schritten partielle Ergebnisse erzeugt und genutzt werden. Auch hierdurch wird ein hohes Maß an Robustheit erreicht: Bei unvorhergesehenem oder fehlerhaftem Input ist es möglich, nicht verarbeitbares Material zu überspringen.

In Fliedner (2007) sind Ergänzungen des PreDS-Parsers u.a. um Heuristiken zur Auflösung unterspezifizierter PreDS-Relationen und um ein Modul zur Anaphernauflösung beschrieben. Beide Erweiterungen wurden erst fertiggestellt, nachdem große Teile des von uns verwendeten Korpus bereits analysiert waren. Sie konnten daher für unsere Arbeiten nicht mehr genutzt werden. Zur Behandlung unterspezifizierter Relationen haben wir einige positionsbasierte Heuristiken eingesetzt (so wurde in Konstellationen mit zwei *NPArgs* generell das nach Oberflächenposition erste als *DSub* und das zweite als *DObj* gewertet, sofern beide Relationen noch nicht vergeben waren). Information zu satzübergreifenden Koreferenzketten, die speziell innerhalb definitorischer Argumentationen ein wichtiges Mittel der Kohärenzerzeugung darstellen (vgl. Kap. 1), stand uns in unserem System leider nicht zur Verfügung. Die von Fliedner entwickelten Komponenten würden mit Sicherheit eine wichtige Ergänzung der hier diskutierten Verarbeitungskette bilden.

**(d) Anpassungen für die Verarbeitung von juristischem Text**

Der PreDS-Parser wurde ursprünglich für die Verarbeitung von Zeitungstext (insbesondere Wirtschaftsmeldungen) entwickelt. Auch die von Fliedner vorgenommene Weiterentwicklung des Parsers[11] betraf vor allem Modifikationen für die Verarbeitung allgemeinsprachlich-journalistischer Sachtexte (der Textgrundlage des von Fliedner entwickelten QA-Systems).

Aufgrund verschiedener Charakteristika der deutschen Rechtssprache (in 4.3 werden wir auf dieses Thema näher eingehen) waren für den Einsatz des PreDS-Parsers in der juristischen Domäne weitere Optimierungen in bestimmten Bereichen nötig, die bei der bisherigen Entwicklung weniger fokussiert wurden. Die für unser Nutzungsszenario am PreDS-Parser vorgenommenen Anpassungen sind vor allem zwei Kategorien zuzuordnen. Zum einen wurden die Grammatik des topologischen Parsers und der PreDS-Konstruktionsprozess systematisch im Hinblick auf definitionsrelevante Konstruktionen ergänzt und verfeinert. Dies betraf insbesondere Kopulakonstruktionen, Satzgefüge mit Relativsätzen und zugehörigen korrelativen Demonstrativa (*... ist derjenige, der...*) bzw. mit freien Relativsätzen (*Zu... zählt, wer...*) sowie schließlich *als*-Prädikative, wie sie z.B. mit Verben des Nennens verwendet werden (*... wird als... bezeichnet*).

Zum anderen war eine umfangreiche Erweiterung der zur *Named Entity*-Erkennung verwendeten Ressourcen erforderlich. Durch diese konnte eine relativ hohe Abdeckung bei der Erkennung von Zitatangaben zu Rechtsquellen und Belegstellen erreicht werden, wie sie für die Verarbeitung von Urteilstexten unerlässlich ist.

Solche Angaben dienen der intertextuellen Verknüpfung zwischen Entscheidungen. Sie können (wie in Kap. 1 und Kap. 2 diskutiert) beispielsweise einer Begriffsbestimmung durch den Verweis auf höchstrichterliche Rechtsprechung besonderes Gewicht verleihen oder, indem sie eine Wiederaufnahme anzeigen, zu ihrer Etablierung als "ständige Rechtsprechung" beitragen. Neben ihrer funktionalen Relevanz müssen Zitatangaben aber auch deswegen mit hoher Zuverlässigkeit verarbeitet werden, weil sie häufig Satzteile darstellen und Fehler bei ihrer Erkennung deswegen den Parsing-Prozesses behindern können.

Generell kann zwischen drei verschiedenen Typen juristischer Zitatangaben unterschieden werden, die in Tabelle 4.2 jeweils mit Beispielen illustriert sind.

Die Beispiele zeigen, dass in keiner der drei Klassen strikt einheitliche Zitierschemata befolgt werden. Bei Normzitaten betreffen die Unterschiede vor allem die Reihenfolge und die genaue Realisierung einzelner Bestandteile (also

[11] Wie bereits angesprochen umfasst Fliedners Arbeit darüber hinaus die Integration weitreichender linguistischer Zusatzfunktionalität bis hin zur Erzeugung framesemantischer Bedeutungsrepräsentationen.

| | |
|---|---|
| Normzitate: | §130 b Satz 2 VwGO |
| | Art. 3 Abs. 1 Satz 2 RuStAÄndG 1974 |
| | Art. 2 Nr. 6 Abs. 2 S. 4 AFWoG NRW |
| | §32 Abs. 1, Ziff. 3, erster Teilstrich StUG |
| | Nr. 1.5.1 der Anlage zur Beihilfeverordnung |
| | Art. 3 der Europäischen Menschenrechtskonvention (EMRK) |
| | UrhG §53 Abs 2 Nr 4 Buchst a |
| Rechtsprechungszitate: | OVG Münster durch Urteil vom 26. April 1977 - XII A 574/74 |
| | VG Köln, Urteil vom 6. Dezember 1977 - 2 K 1001/71 |
| | BGHZ 69, 1, 22 ff. |
| | ThürOVG, B. v. 9.8.1996 - 2 EO 669/96 - NVwZ - RR 1997, 287 = DVBl. 1996, 1446 = NJ 1997, 102 = ThürVBl. 1997, 34 |
| | BGH, 1976-12-22, III ZR 62/74, BGHZ 69, 1 |
| | Parallelsache III ZR 89/75 (Urteil ebenfalls vom 22. Dezember 1976 = LM BGB § 823 (Ad) Nr. 10, dort zu II 3 b) |
| | ebenso : OLG Hamm , Urteil vom 18. 12. 1980 - 5 U 87 / 80 - , AgrarR 1981 , 288 m. Anm. Bendel |
| Literaturzitate: | Seibert, DVBl. 1997, 932, 934 |
| | Buchholz 402.44 § 15 VersG Nr. 6 |
| | Ott/Wächtler, VersG, 6. Aufl. 1996, § 1 Rdn. 6 |
| | Maunz/Dürig-Herzog, GG, Art. 8 Rdn. 136 |
| | Friauf in: v. Münch, Bes.VerwR 6. Aufl. S. 597 |
| | BGB-RGRK 12. Aufl. § 839 Rdn. 583 m.w.Nachw. |
| | Scheerbarth, Das allgemeine Bauordnungsrecht, 2. Aufl. S. 361 |

Tabelle 4.2: Beispiele für verschiedene Typen juristischer Zitatangaben

etwa die Verwendung von Lang- oder Kurztiteln in Kombination mit vollen Bezeichnungen oder Abkürzungen für Gliederungspunkte). Bei Rechtsprechungs- und Literaturzitaten ist zudem die Art der enthaltenen Angaben selber unterschiedlich. In allen drei Klassen schließlich bestehen vielfältige Möglichkeiten der Koordination von Angaben auf allen Ebenen, und es können in begrenztem Maß natürlich-sprachliche Bestandteile integriert werden. Insbesondere kann so bei Rechtsprechungszitaten durch Anmerkungen wie *ebenso* oder *anders* angedeutet werden, ob die zitierte Entscheidung eine Ansicht teilt oder ablehnt.

Die *Named Entity*-Erkennungskomponente der in Fliedner (2007) beschriebenen Version des PreDS-Parsers umfasst Regeln zur Verarbeitung der gängigsten Typen von Normzitaten. Diese haben wir stark ausgeweitet. Weiterhin haben wir Teilgrammatiken zur Erkennung von Rechtsprechungszitaten unter Angabe von Gericht und Datum (vgl. die ersten zwei Beispiele an entsprechender Stelle in Tabelle 4.2) sowie Rechtsprechungszitaten aus der Zeitschriftenliteratur (vgl. das dritte Beispiel) hinzugefügt. Die Verarbeitung genereller Literaturzitate erwies sich als problematisch, da Titel von Publikationen keinen verallgemeinerbaren Längen- oder Strukturbeschränkungen unterliegen. Unser System enthält keine eigenen Regeln zur Erkennung solcher Zitate. Sie können allerdings dennoch verarbeitet werden, soweit sie – wie einige der Beispiele in Tabelle 4.2 – Personennamen mit durch die vorhandenen Regeln abgedeckten Publikationstiteln und Stellenangaben verbinden.

Als Quelle für die Anker in den Regeln zur Zitaterkennung dienen eigene Gazeteers mit Voll- und Kurztiteln von Gesetzen und Verordnungen (etwa 12 000 Einträge) sowie juristischen Periodika (etwa 1500 Einträge) und Bezeichnungen deutscher und der wichtigsten internationalen Gerichte (etwa 150 Einträge). Diese haben wir aus verschiedenen frei verfügbaren Quellen (hauptsächlich Webseiten des Bundes und der Bundesländer, verschiedener Anbieter von Entscheidungssammlungen und juristischer Verlage) zusammengestellt. Die zugehörigen Grammatikregeln beschreiben für Normzitate in erster Linie unterschiedliche Gliederungsschemata und Koordinationsmöglichkeiten. Für Rechtsprechungszitate aus Zeitschriften decken sie die Kombinations- und Ausdrucksmöglichkeiten aus Jahrgangs-, Band- und Seitenangaben ab. Für Rechtsprechungszitate nach Gericht kodieren sie unterschiedliche Schemata zur Angabe von Spruchkörper, Gerichtsort und Aktenzeichen sowie Kombinationen mit Datumsangaben.

#### (e) Evaluation

Um die Auswirkungen der von uns in den PreDS-Parser integrierten Anpassungen zu erfassen, haben wir diese einzeln und in Kombination gegen die in Fliedner (2007) verwendete Version des Parsers evaluiert. Im Folgenden verwenden wir die Abkürzungen PreDS.GF und PreDS.CORTE zur Bezeichnung der Versionen des PreDS-Parsers aus Fliedner (2007) bzw. mit den von uns für das CORTE-System vorgenommenen Anpassungen.

**Zitaterkennung.** In Tabelle 4.3 ist die Erkennungsleistung der Zitatgrammatik in PreDS.GF und der von uns erweiterten Grammatikversion verglichen.[12]

[12] Da Literaturzitate von der *Named Entity*-Grammatik des PreDS-Parsers nicht systematisch erfasst werden, haben wir sie für die Evaluation außer Betracht gelassen. Der Häufigkeit nach

| | Normzitate | | Rechtsprechungszitate | |
|---|---|---|---|---|
| | *alt* | *neu* | *alt* | *neu* |
| **Erkennungsgenauigkeit** | | | | |
| *Treffer insgesamt* | 58 | 125 | - | 32 |
| davon: | | | | |
| *korrekt* | 20 (≈ 34%) | 62 (≈ 50%) | - | 9 (≈ 28%) |
| *partiell* | 32 | 19 | - | 22 |
| *inkorrekt* | 6 | 44 | - | 1 |
| **Erkennungsrate** | | | | |
| *Annotierte Quellenangaben igs.* | 65 | 65 | - | 32 |
| davon: | | | | |
| *gefunden* | 20 (≈ 31%) | 51 (≈ 78%) | - | 6 (≈ 19%) |
| *partiell gefunden* | 23 | 9 | - | 18 |
| *nicht gefunden* | 22 | 5 | - | 8 |

Tabelle 4.3: Qualität der Erkennungsergebnisse für Norm- und Rechtsprechungszitate

Für die Ermittlung dieser Werte wurden in zwei zufällig ausgewählten Entscheidungen aus dem CORTE-Korpus (LSG Essen 17. Senat vom 8. Oktober 1997, L 17 U 136/96 und BVerwG 4. Senat vom 29. Oktober 1982, 4 C 51/79) manuell sämtliche Zitate aus Normen und Rechtsprechung markiert. Die insgesamt 338 Sätze dieser Dokumente enthielten 65 Quellenangaben zu Normzitaten und 32 Quellenangaben zu Rechtsprechungszitaten. Koordinationen mehrerer Quellenangaben wurden als eine Einheit annotiert.

Natürlichsprachliche Ergänzungen (etwa Konjunktionen oder Kommentare) innerhalb von Quellenangaben mussten nicht erkannt werden. Treffer wurden bei der Berechnung der Erkennungsgenauigkeit als *korrekt* gewertet, wenn sie in einer annotierten Quellenangabe die eigentliche Quelleninformation alleine oder in Kombination mit anderen Treffern vollständig abdeckten. Da somit auf eine Quellenangabe mehrere als korrekt gewertete Treffer kommen konnten, liegt die Anzahl der korrekten Treffer im Fall der erweiterten Zitaterkennung höher als die Zahl der annotierten Quellenangaben. Aus demselben Grund besteht auch eine Differenz zwischen den Werten für partiell erkannte Quellenangaben bei Erkennungsgenauigkeit und -rate.

scheint diese Gruppe vergleichbar mit Rechtsprechungszitaten zu sein. Insgesamt fanden sich in den zwei aufgearbeiteten Entscheidungstexten 42 solche Zitate. Eine entsprechende Erweiterung der *Named Entity*-Grammatik erscheint daher als sinnvolle Ergänzung, um eine weitere Verbesserung der Parseergebnisse des PreDS-Parsers zu erreichen.

Für Normzitate zeigt Tabelle 4.3 eine deutliche Verbesserung sowohl der Erkennungsgenauigkeit (*Präzision*) als auch der Erkennungsrate (*Recall*) gegenüber PreDS.GF. Die Erkennungsgenauigkeit von knapp 50% erscheint dennoch zunächst als unzureichend. Gleiches gilt für die gesamte Erkennungsleistung bei Rechtsprechungszitaten. Bei der Interpretation der in Tabelle 4.3 angegebenen Werte muss jedoch beachtet werden, dass nicht alle partiellen oder inkorrekten Ergebnisse in gleichem Maße problematisch sind.

So handelt es sich im Falle der Normzitate bei 17 der 44 inkorrekten Treffer der erweiterten Zitaterkennung um Abkürzungen oder Nummerierungen, die zwar keine Normstellen bezeichnen, aber auch nicht in anderer Weise durch den PreDS-Parser verarbeitbar wären. Die fälschliche Erkennung dieser Elemente als Zitatangaben führt somit zumindest nicht zu einer Verschlechterung der Gesamtanalyse des PreDS-Parsers.

Im Fall der Rechtsprechungszitate enthält zudem in einem Großteil der als partiell gezählten Analysen (18 für die Erkennungsgenauigkeit und 15 für die Erkennungsrate) der erkannte Teil der Quellenangabe noch genug Informationen, um die Quelle zuverlässig zu identifizieren. Die Analyse ist also für eine etwaige Rückverfolgung des Zitats hinreichend. Der Hauptgrund für den hohen Anteil solcher Fälle liegt in der bereits angesprochenen Vielfalt teilweise idiosynkratisch verwendeter koordinierender und ergänzender Ausdrücke. Berücksichtigt man auch diese Treffer entsprechend (d.h., lässt man sie jeweils außer acht bzw. zählt sie als korrekt) ergibt sich eine Erkennungsrate von 66% und eine Genauigkeit von 84%.

**Gesamtperformanz.** Wir evaluieren nun die Gesamtperformanz von PreDS.CORTE im Vergleich zu PreDS.GF und ermitteln dabei den Beitrag der einzelnen integrierten Modifikationen im Kontext des Gesamtsystems. Dabei legen wir eine rechtssprachliche Korpus-Stichprobe zu Grunde, die auch schon von Fliedner (2007) als Evaluationsgrundlage verwendet wurde. Sie umfasst 100 Sätze aus zwei verwaltungsrechtlichen Entscheidungen (Verwaltungsgerichtshof Baden-Württemberg 8. Senat vom 17. September 1999, 8 S 2042/99 und VG Karlsruhe 14. Kammer vom 9. August 1999, 14 K 1009/99). Die Annotation erfolgte nach dem unter anderem in Briscoe und Carroll (2000) beschriebenen *grammatical relations*-Annotationsschema. Nach diesem Schema werden Sätze mit flachen Kollektionen von Dependenz-Tupeln annotiert, ohne dass dabei lokale Substrukturen berücksichtigt werden. Die Annotation lässt somit eine robuste globale Evaluation der Parser-Performanz zu. Eine genauere Bewertung der Ergebnisqualität (etwa für bestimmte Konstruktionen) erfordert allerdings weitere Untersuchungen. Vgl. zu einer detaillierteren Dis-

kussion des Annotationsschemas und der Evaluationsmethodologie Fliedner (2007), 350ff.

Die Annotation erfolgte zunächst doppelt (durch zwei Computerlinguistikstudenten). Die doppelte Annotation wurde dann zu einem Goldstandard zusammengeführt. Sie stützte sich auf die in Fliedner und Braun (2005) festgehaltenen Richtlinien. Für die rechtssprachliche Korpus-Stichprobe wurde eine Übereinstimmung der Annotatoren von 94% erreicht.

Für die Evaluation der Analyseergebnisse des PreDS-Parsers wurden diese ebenfalls in das *grammatical relations*-Format übersetzt. Sowohl der Übersetzungsvorgang als auch der anschließende Vergleich mit dem Goldstandard lassen dabei verschiedene Parametrisierungen zu, die jeweils mit Unterschieden im ermittelten Übereinstimmungsgrad einhergehen. Die Parametrisierungen resultieren aus der Möglichkeit, Anknüpfungspunkte und Dependenzrelationen in PreDS unterspezifiziert zu repräsentieren. Fliedner definiert drei verschiedene Szenarios, in denen (1) exakte Übereinstimmung zwischen Goldstandard und Parseergebnis gefordert wird (*strict*), (2) Unterschiede in der Einordnung von Dependenzen als Argument oder Modifikation zugelassen werden (*relaxed*) bzw. (3) zusätzlich angenommen wird, dass die unterspezifizierten Dependenzen und Anknüpfungspunkte soweit wie möglich entsprechend dem Goldstandard resolviert wurden (*lenient*). Wir geben in Tabelle 4.4 die von Fliedner für das rechtssprachliche Subkorpus berichteten Übereinstimmungswerte in den Szenarios *strict* und *lenient* wieder und stellen diesen die entsprechenden Werte für PreDS.CORTE in drei Varianten gegenüber: (a) mit den oben genannten Anpassungen der Grammatik, jedoch ohne die erweiterte Zitaterkennung, (b) mit unmodifizierter Grammatik, jedoch mit erweiterter Zitaterkennung und (c) mit allen Anpassungen.

Von den insgesamt annotierten 100 Sätzen konnte für 91 eine PreDS ermittelt werden. Dies entspricht in etwa der auch auf dem Gesamtkorpus erzielten Rate von ca. 88% parsbaren Sätzen. Für die Berechnung der in Tabelle 4.4 angegebenen Scores wurden für jeden Satz sämtliche annotierten Relationen gefordert (eine Betrachtung lokaler Strukturen lässt das Annotationsformat wie oben angesprochen nicht zu). Die Zahlen beinhalten also den gesamten Recall-Verlust durch vollständig unanalysierbare Sätze, auch wenn dieser teilweise eher durch Schwierigkeiten bei der Vorverarbeitung als durch Probleme des Parsers selber bedingt sein dürfte.[13]

Die Anpassung der Grammatik führt unter beiden betrachteten Bedingungen zu einer deutlichen Verbesserung sowohl der Präzision als auch des Recall der Parseergebnisse. Die erweiterte Zitaterkennung alleine bewirkt zwar eine Er-

[13] Insbesondere unübliche Nummerierungen, ungewöhnlich markierte Parenthesen sowie Erfassungsfehler wie Blockschrift oder gesperrte Eingabe können oft nicht vollständig normalisiert werden.

| | *PreDS.GF* | *angepasste Grammatik* | *verbesserte Zitaterkennung* | *alle Anpassungen* |
|---|---|---|---|---|
| **strict** | | | | |
| *Präzision* (%) | 46,15 | 49,33 | 49,57 | 49,42 |
| *Recall* (%) | 40,46 | 44,15 | 37,68 | 44,05 |
| *f-Score* (%) | 43,12 | 46,60 | 42,82 | 46,58 |
| **lenient** | | | | |
| *Präzision* (%) | 61,58 | 62,75 | 64,11 | 63,61 |
| *Recall* (%) | 56,12 | 58,62 | 51,01 | 58,84 |
| *f-Score* (%) | 58,72 | 60,61 | 56,81 | 61,13 |

Tabelle 4.4: Qualität der Analyseergebnisse des PreDS-Parser für juristischen Text vor und nach der Anpassung

höhung der Präzision, jedoch eine deutliche Einschränkung des Recall. Dies deutet darauf hin, dass zum Teil Material fälschlicherweise in Zitatangaben integriert wird und somit nicht mehr für den Aufbau grammatischer Relationen zur Verfügung steht. Teils dürfte der Effekt aber auch darauf beruhen, dass in manchen Fällen im Goldstandard *grammatical relations* auch innerhalb strukturierter Zitatangaben annotiert wurden.

Die Kombination aller Anpassungen führt in sämtlichen Fällen zu einer deutlichen Verbesserung gegenüber der unmodifizierten Version des Parsers. Im Szenario *strict* kann in der Kombination die Recall-Verschlechterung durch die angepasste Zitaterkennung insgesamt aufgewogen werden. Im Szenario *lenient* wird durch die Kombination unserer Optimierungen sogar ein gegenüber der Grammatikanpassung allein noch leicht verbessertes Gesamtergebnis erzielt.

## 4.3 Datengrundlage

Neben den verfügbaren sprachtechnologischen Werkzeugen sind die sprachlichen Besonderheiten der Datengrundlage ein weiterer wesentlicher Faktor beim Design eines Informationszugriffssystems, das sich auf eine computerlinguistische Verarbeitungskette stützt.

Für die in dieser Arbeit beschriebenen Untersuchungen hatten wir Zugriff auf eine umfangreiche Auswahl von Dokumenten aus der von *juris* geführten Entscheidungssammlung. Bereits für die in Kap. 1 und Kap. 2 beschriebenen Studien haben wir uns auf einen Ausschnitt aus diesem Datenbestand gestützt

(vgl. auch Anhang A). Im Rest dieses Kapitels werden wir das gesamte von uns genutzte Korpus genauer beschreiben. Zwar haben wir diese Daten in erster Linie als Grundlage für die Entwicklung und Evaluation unseres Definitionsextraktionssystems verwendet. Das Korpus stellt jedoch auch aus linguistischer Sicht eine interessante Ressource dar, denn es ermöglicht die Überprüfung von Hypothesen über die Rechtssprache als Fachsprache. Dies gilt besonders aufgrund der oben beschriebenen sprachtechnologischen Aufarbeitung.

Wir gehen daher im Folgenden nach Angaben zum Aufbau des Korpus in einem kurzen Exkurs auf einige statistische Befunde zu dessen "typisch rechstssprachlichen" Eigenschaften ein. Die meisten der dargestellten Beobachtungen stützen dabei auch das zuvor über die besonderen Anforderungen bei der automatischen Verarbeitung von Rechtstexten Gesagte.

### 4.3.1 Korpus

Insgesamt sind über die Recherchedienste der Firma *juris* knapp eine Million Entscheidungstexte zugänglich (vgl. Kap. 3). Der Ausschnitt dieses Datenbestandes, den wir für unsere Arbeiten genutzt haben, umfasst (neben den in den vorigen Kapiteln angesprochenen manuell aufgearbeiteten 100 Urteilstexten) insgesamt 33 579 weitere Entscheidungen. Wir bezeichnen diese im Folgenden als das *CORTE-Korpus* oder auch *CORTE-Großkorpus*. Um eine für unsere Zwecke möglichst relevante Datengrundlage zu erhalten, haben wir das CORTE-Korpus unter drei Gesichtspunkten zusammengestellt:

- Es sollte einen "überschaubaren" Rechtsbereich als Ganzen möglichst vollständig abdecken (wir haben uns für das Umweltrecht entschieden).
- Es sollte für einen nach dem jährlichen Aufkommen besonders relevanten Bereich (hier haben wir das Verwaltungsrecht gewählt) eine repräsentative Abdeckung aufweisen.
- Es sollte die Nutzung des *juris*-Definitionsregisters (vgl. Kap. 2) als Referenz bei unseren Arbeiten ermöglichen.

Entsprechend ist das Gesamtkorpus in drei Subkorpora unterteilt. Die Verteilung der Entscheidungen auf diese Subkorpora ist in Tabelle 4.5 angegeben. Das erste umfasst die von *juris* dokumentierten umweltrechtlichen Urteile aus den Jahren 1979 bis 2001 und das zweite die Verwaltungsrecht-Jahrgänge 1995–1997. Das dritte Subkorpus enthält einen großen Teil (etwa 80%) der Urteile, für die im *juris*-Definitionsregister Eintragungen vorhanden sind (siehe 2.4.2).

| **Gesamtwerte** | *Sätze* | *davon in Leitsätzen und Gründen* | *Tokens* | *davon in Leitsätzen und Gründen* |
|---|---|---|---|---|
| | 2 307 189 | 1 645 799 | 71 409 021 | 54 736 111 |

| **Aufteilung** *Subkorpus* | *Anzahl* | *kurzerfasst (< 10 Sätze)* | *∅ Gesamtlänge (Sätze)* | *davon in Leitsätzen und Gründen* |
|---|---|---|---|---|
| *UWR 1979–2001* | 9068 | 4310 | 139 | 83 |
| *VWR 1995–1997* | 16 239 | 7501 | 109 | 60 |
| *Definitionsregister* | 8272 | 62 | 107 | 75 |
| *Gesamtkorpus* | 33 579 | 11 873 | 115 | 71 |

Tabelle 4.5: Aufbau des CORTE-Korpus

Tabelle 4.5 sind außerdem Angaben zur Länge der Dokumente in den einzelnen Subkorpora zu entnehmen. Die umwelt- und verwaltungsrechtlichen Urteile sind zum Teil nur in kurzen Auszügen erfasst. Um eine Verzerrung durch diese Fälle auszuschließen, haben wir für die Berechnung der Dokumentlängen in Tabelle 4.5 Dokumente mit einer Länge von bis zu zehn Sätzen außer Betracht gelassen. Das längste Dokument umfasst zwar 1577 Sätze, jedoch sind nur etwa 6% der Dokumente länger als 200 Sätze. Die Standardabweichung der Gesamtlänge liegt bei 83 Sätzen. Durchschnittlich etwa 60% der Sätze eines Dokuments liegen in den für die Definitionssuche relevanten begründungsbezogenen Entscheidungsteilen.

In Tabelle 4.6 ist der Aufbau des Gesamtkorpus nach weiteren Kriterien (Jahrgang, Gerichtsbarkeit und Instanz sowie – soweit diese Information für uns verfügbar war – der *juris*-Sachgebietszuordnung) genauer aufgeschlüsselt. Das Korpus umfasst schwerpunktmäßig obergerichtliche Entscheidungen aus den 1980er und 1990er Jahren. Durch die vollständige Aufnahme der verwaltungsrechtlichen Rechtsprechung der Jahrgänge 1995–1997 ergibt sich natürlich zum einen ein hoher Anteil an Entscheidungen der Verwaltungsgerichtsbarkeit und zum andern eine besonders hohe Abdeckung für diesen Zeitraum. Wie die Aufschlüsselung nach Sachgebieten erkennen lässt, repräsentieren die Texte im Korpus allerdings insgesamt dennoch eine Vielzahl verschiedener Themen und Rechtsbereiche.

**Jahrgänge**

| *Jahr* | *Definitions-register* | *Umwelt-recht* | *Verwaltungs-recht* |
|---|---|---|---|
| 2006 | 38 | | |
| 2005 | 345 | | |
| 2004 | 378 | | |
| 2003 | 385 | | |
| 2002 | 298 | | |
| 2001 | 350 | 402 | |
| 2000 | 396 | 492 | |
| 1999 | 405 | 476 | |
| 1998 | 361 | 538 | |
| 1997 | 293 | 585 | 5329 |
| 1996 | 197 | 543 | 5394 |
| 1995 | 242 | 537 | 5516 |
| 1990–94 | 1123 | 2495 | |
| 1985–89 | 771 | 1797 | |
| 1980–84 | 996 | 1204 | |
| 1970–79 | 1416 | | |
| vor 1970 | 278 | | |

**Sachgebiete**[14]

| *Sachgebiet* | *Anzahl* |
|---|---|
| Allgemeines Verwaltungsrecht | 704 |
| Arbeitsrecht | 508 |
| Besonderes Verwaltungsrecht | 2889 |
| Bürgerliches Recht | 2823 |
| Europarecht | 70 |
| Gerichtsverfassung und -zuständigkeit | 525 |
| Handels- und Wirtschaftsrecht | 2041 |
| Recht des öffentlichen Dienstes | 374 |
| Regionen, Rechtswissenschaft, Kirchenrecht | 1577 |
| Sozialrecht | 1508 |
| Staats- und Verfassungsrecht | 752 |
| Steuerrecht | 2291 |
| Strafrecht | 1386 |
| Völkerrecht | 1 |
| Weitere Gerichtsverfahrensordnungen | 1969 |
| Zivilprozessrecht | 1023 |

**Gerichtsbarkeit und Instanz**[15]

| | *Bundes-obergericht* | *Oberstes Gericht (Land)* | *Mittleres Gericht (Land)* | *Unterstes Gericht (Land)* |
|---|---|---|---|---|
| Zivilgericht | | 7 | 6 | |
| Strafgericht | | 3 | 3 | |
| Verwaltung | 4077 | 10 137 | 3618 | |
| Sozial | 1299 | 900 | 150 | |
| Arbeits | 766 | 960 | 52 | |
| Finanz | 496 | | 492 | |
| Verfassung | 588 | 125 | | |
| ordentl. Gerichtsbarkeit | 3090 | 1358 | 4498 | 443 |

Tabelle 4.6: Aufbau des CORTE-Korpus nach Jahrgang, Gerichtsbarkeit und Sachgebieten

[14] Insgesamt lagen uns nur für 10 301 Urteile Angaben zu Sachgebieten vor. Die meisten Urteile sind mehreren Sachgebieten zugeordnet.

[15] Bei den in der Aufstellung fehlenden Urteilen handelt es sich größtenteils um Entscheidungen von Gerichten der Berufsgruppen, ausländischen und internationalen Gerichten.

### 4.3.2 Sprachliche Besonderheiten

Als zentrales Charakteristikum der Rechtssprache gilt im allgemeinen ihre hohe Komplexität. Aus dieser sprachlichen Komplexität resultiert mangelnde Rechtsverständlichkeit. Sie schränkt den Zugang des Bürgers zum Recht ein und steht somit in direktem Widerspruch zu rechtsstaatlichen Grundprinzipien. Dieses Phänomen wurde daher bereits vielfach und aus den verschiedensten Perspektiven untersucht (vgl. aktuell etwa Eichhoff-Cyrus und Antos (2008) und Klein (2004)).

Die dabei ausgemachten Gründe liegen teils in allgemeinen Faktoren, wie etwa der Komplexität der in Rechtstexten behandelten Materie und der starken Institutionalisierung der Rechtssprache, die sich vor allem in der ständigen impliziten und expliziten Bezugnahme auf andere Rechtstexte und umfassendes fachliches Hintergrundwissen äußert (vgl. Busse (1992)). Daneben wird aber regelmäßig auch eine Anzahl im engeren Sinne sprachlicher Faktoren benannt. Hierzu zählt insbesondere die Tendenz zu einem "Nominalstil" und sehr umfangreichen Satzgefügen. Diese Eigenschaften werden zwar an den verschiedensten Stellen (neben linguistischen Studien etwa in Stilleitfäden für Juristen und auch in der öffentlichen Diskussion) immer wieder durch Beispiele belegt. Sie sind allerdings bisher erst relativ selten korpusbasiert untersucht worden.

#### (a) Bisherige Untersuchungen

In jüngerer Zeit haben sich mehrere empirische Studien mit Rechtstexten befasst, dabei standen jedoch psycholinguistische Fragen im Vordergrund. Das Projekt *Empirische Studie zur Verständlichkeit von Versicherungsklauseln* der Arbeitsgruppe *Sprache des Rechts* der Berlin-Brandenburgischen Akademie der Wissenschaften untersucht die Verständlichkeit von Versicherungsbedingungen. Der dafür gewählte Zugang ist ein direkt experimenteller, und das Projekt befasst sich nicht näher mit einzelnen syntaktischen oder lexikalischen Merkmalen der Textgrundlage (Dietrich und Schmidt (2002); Schmidt (2001); Dietrich (2000)). Dagegen beschreiben Hansen u. a. (2006) und Neumann (2009) Untersuchungen zur Verständlichkeitsoptimierung, die u.a. auf einem etwa 35 600 Textwörter umfassenden Korpus aus Entscheidungen des Bundesverfassungsgerichts basieren. Sie erheben für dieses auch Daten zur Häufigkeit von Nominalisierungen sowie zur syntaktischen Komplexität. Außerdem nehmen sie einen Vergleich mit Zeitungstexten vor (Zeitungsberichte über die Entscheidungen in ihrem Korpus).

Die ausführlichste Untersuchung zu den angesprochenen Phänomenen stammt jedoch bereits aus den 1970er-Jahren und liegt mit Wagner (1972) vor. Wagner befasst sich mit der deutschen Verwaltungssprache im allgemeinen

(die viele Gemeinsamkeiten mit der Sprache in Urteilstexten aufweist) anhand einer über die vier Textsorten *Verwaltungsakte*, *Verwaltungsvorschriften*, *Schriftverkehr* und *informative Schriften* balancierten Auswahl von rund 1000 Einzelsätzen.

Mit ihren Ergebnissen belegt Wagner die Bedeutung des Nominalstils in der Verwaltungssprache. Diesen betrachtet sie als generell funktional begründet. Er dient nach ihrer Analyse zum einen der Informationsverdichtung – statt als komplette Nebensätzen werden Inhalte als nominale Konstruktionen im Satz integriert. Ein extremes Beispiel aus unseren Korpora ist in (4.7) angeführt.

(4.7) Insoweit setzt die Feststellung einer Verunstaltung kein so krasses geschmackliches Unwerturteil wie das einer das ästhetische Empfinden verletzenden Häßlichkeit voraus.

(Oberverwaltungsgericht Berlin 2. Senat, 31. Juli 1992, AZ 2 B 14.90, juris)

Zum anderen erlaubt der Nominalstil eine sprachliche Vereinheitlichung, indem Prädikate in einen festen nominalen Bestandteil und ein veränderliches semantisch schwaches Verb aufgespalten werden (z.B. *Voraussetzung haben* für *voraussetzen*).

Als weiteres Kennzeichen der Verwaltungssprache stellt Wagner die unpersönliche Ausdrucksweise heraus. Auch in diesem Zusammenhang kommt ersichtlich dem Gebrauch von Nominalisierungen eine (von Wagner allerdings nicht erwähnte) Funktion zu: Im Unterschied zu satzförmigen Realisierungen propositionaler Inhalte können bei nominaler Ausdrucksweise die Handlungsbeteiligten unbenannt bleiben.

Dagegen findet Wagner die der Verwaltungssprache oft vorgeworfene Tendenz zu “Schachtelsätzen” durch ihre Daten nicht bestätigt. Sie zieht als Vergleich Untersuchungen zu den Genres Zeitungs- und Sachbuchtext heran. Sowohl die Satzlängen als auch die Einbettungstiefen der Satzgefüge in ihrem verwaltungssprachlichen Korpus liegen unter den Werten für die Vergleichsdaten. Dies deutet sie zum einen als Effekt der Informationsverdichtung durch Nominalisierung, zum anderen auch als Anzeichen für ein einsetzendes Bemühen öffentlicher Institutionen um eine stilistisch bessere Ausdrucksweise.

#### (b) Ergebnisse aus dem CORTE-Großkorpus

Wir gehen nun auf die im CORTE-Großkorpus ermittelten Daten zu den angesprochenen Formulierungsmerkmalen sowie zu einigen weiteren statistisch erfassbaren sprachlichen Charakteristika ein. Diese vergleichen wir mit den Ergebnissen aus den Studien von Wagner und Neumann. Zusätzlich ziehen wir

entsprechende Werte für das ECI-FR Korpus heran, das Zeitungstext aus einem Jahrgang der *Frankfurter Rundschau* enthält (ca. 41 Millionen Textwörter).

Wir sind uns bewusst, dass als Vorbedingung für weitergehende Schlussfolgerungen aus unseren Ergebnissen noch eine Reihe methodologischer Fragen zu klären sind. Dies betrifft zunächst einmal die Frage der Eignung der zwei verwendeten Korpora. Wir haben das CORTE-Korpus nicht in erster Linie im Hinblick auf die sprachliche Repräsentativität der einbezogenen Texte, z.B. hinsichtlich stilistischer Unterschiede zwischen Rechtsgebieten, zusammengestellt. Auch das ECI-FR-Korpus deckt mit Texten aus der Frankfurter Rundschau nur eine recht spezielle Auswahl aus dem Genre Zeitungstext ab. Dies zeigt sich beispielsweise daran, dass das Wort *Frankfurter* das neunthäufigste Adjektiv im Korpus ist. Weiterhin ist davon auszugehen, dass die beiden Korpora neben dem Genreunterschied hinsichtlich einer Reihe weiterer, von uns nicht kontrollierter Parameter differieren. So ist es vor allem im lexikalischen Bereich nur schwer abzuschätzen, ob Differenzen zwischen den Korpora tatsächlich genrebedingt sind, oder eher Themen- und Domänenunterschiede reflektieren. Schließlich wäre auch die statistische Signifikanz der festgestellten Befunde noch zu überprüfen.

Eine entsprechende Ausarbeitung der Untersuchung werden wir im Rahmen des gegenwärtigen Exkurses nicht leisten. Unsere Ergebnisse stehen daher unter der angesprochenen methodologischen Kautel uns können nur als indikativ betrachtet werden. Wir haben uns aber entschlossen, sie hier dennoch wiederzugeben. Unseres Wissens bestand bisher überhaupt nicht die Möglichkeit, die von der Fachsprachenforschung identifizierten linguistischen Charakteristika der Rechtssprache empirisch auf einer Datenbasis zu untersuchen, die in Umfang und Aufarbeitung mit dem CORTE-Korpus vergleichbar wäre.

Wir betrachten zunächst die lexikalische Ebene und gehen speziell auf Daten zum Auftreten von Nominalisierungen ein. Dann untersuchen wir verschiedene syntaktische Charakteristika. Für Angaben zu Lemmata und POS stützen wir uns auf eine durch den in 4.2.2 beschriebenen TreeTagger (Schmid (1994)) vorverarbeitete Version der Korpora. Um auch über syntaktische Strukturen Vergleiche mit dem CORTE-Korpus anstellen zu können, haben wir zudem einen ca. 900 000 Textwörter umfassenden Ausschnitt des FR-Korpus mittels des PreDS-Parsers analysiert (den auch im TIGER-Korpus, das in Brants u. a. (2002) beschrieben ist, enthaltenen Teil).

**Lexis.** In Tabelle 4.7 sind einige Daten zum Aufbau der beiden Korpora auf lexikalischer Ebene zusammengefasst. Auffallend ist zunächst einmal die in den Korpora deutlich unterschiedliche *Type / Token-Relation*, also das Verhält-

| | *Urteilstexte* | *Zeitungstext* |
|---|---|---|
| **Gesamtzahlen** | | |
| *Textwörter* | 71 409 021 | 40 856 149 |
| *Lemmata* | 777 644 | 514 436 |
| *Type / Token-Relation (gemittelt)* | 3,12% | 7,34% |
| **Häufigkeit der Wortklassen (% der Textwörter)** | | |
| *Inhaltswörter* | 43,50 | 46,03 |
| *Präposition* | 9,72 | 9,20 |
| *Konjunktion* | 2,15 | 2,66 |
| *Subjunktion* | 1,43 | 0,86 |
| *Partikel* | 1,64 | 0,98 |
| *Artikel* | 12,47 | 9,50 |
| *Pronomen* | 5,12 | 5,40 |
| *Modal- bzw. Hilfsverb* | 4,25 | 3,80 |
| *Sonstiges* | 19,70 | 21,57 |

Tabelle 4.7: Korpusgröße und Häufigkeit der Wortklassen (nach Textwörtern)

nis von Lemmata zu Auftreten.[16] Der Wortschatz des FR-Korpus ist demnach erheblich umfangreicher als der des CORTE-Korpus. Dieser Unterschied geht zum Teil darauf zurück, dass im FR-Korpus eine viel größere Anzahl von Eigennamen als im CORTE-Korpus enthalten ist (die Anzahl liegt etwa fünfmal höher). Diese haben oft nur sehr wenige Auftreten. Allerdings liegt der Umfang des Wortschatzes für die Inhaltswörter, die wir in Tabelle 4.7 als eine Kategorie zusammengefasst haben, auch ansonsten grob zweimal so hoch wie im CORTE-Korpus. Es erscheint plausibel, dies als einen Effekt der Verwendung schematischer Formulierungsmuster und normierter Terminologie in der Rechtssprache zu deuten, im Gegensatz zum Streben nach Ausdrucksvielfalt beim journalistischen Schreiben. Erhärten ließe sich diese Vermutung allerdings nur durch eine detailliertere Analyse, denn im Bereich der Inhaltswörter fallen domänen-

bedingte Effekte verglichen mit textsortenbedingten Einflüssen besonders stark ins Gewicht.

Auch im Bereich der Funktionswörter bestehen deutliche Unterschiede zwischen den beiden Korpora. Das juristische Korpus weist einen höheren Anteil an Subjunktionen und einen niedrigeren Anteil an Konjunktionen auf als das FR-Korpus. Dies deutet auf einen hypotaktischen Stil und somit eine erhöhte Satzkomplexität hin. Die relativ große Zahl an Partikeln dürfte vor allem im Zusammenhang mit der differenzierten argumentativen Funktion vieler Sätze in Urteilsbegründungen stehen, denn diese wird in den meisten Fällen durch Kombinationen von Negation und Fokuspartikeln markiert (vgl. Kap. 2). Bei den Hilfs- und Modalverben geht die Differenz zwischen beiden Korpora vor allem auf eine höhere Zahl von Auftreten der Verben *sein* und *haben* im juristischen Korpus (38% bzw. 17% mehr als im Zeitungstext) zurück, die durch die Gebräuchlichkeit der modalen Umschreibungen *hat zu* und *ist zu* erklärbar ist.

Der (auch im Verhältnis zur Anzahl der Nomina) deutlich erhöhte Anteil von Artikeln läßt eine größere Häufigkeit des Nullartikels in Zeitungstexten vermuten. Hierfür kommen verschiedene Gründe in Betracht. Der Nullartikel steht im Deutschen unter anderem bei Abstrakta, Stoffbezeichnungen und generell bei indefiniten Plural-Nominalphrasen. Besonders plausible Erklärungen scheinen uns hier jedoch der abkürzende Stil in Schlagzeilen und eine Anzahl von Einzeleffekten zu sein. So tritt das häufigste Substantiv im FR Korpus, *Uhr*, bei Uhrzeitangaben (wie etwa *um 17:30 Uhr*) ohne Artikel auf. Dagegen werden die in Urteilstexten sehr häufigen Rollenbezeichnungen *Kläger(in)* und *Beklagte(r)* dort meist mit definitem Artikel zur anonymisierten Referenz auf die Prozessbeteiligten gebraucht.

**Nominalisierungen.** Schon der erste Eindruck bei der Beschäftigung mit Rechtstexten lässt kaum Zweifel daran, dass die Verwendung aller Arten von Nominalisierungen in der Rechtssprache stilprägend ist. Die in Tabelle 4.8 angeführten Zahlen liefern eine quantitative Bestätigung für diesen Eindruck.

Insgesamt machen die hier untersuchten Nominalisierungen im juristischen Korpus mehr als ein Viertel aller nominalen Tokens aus. Die Häufigkeit sowohl regulärer als auch derivationaler Nominalisierungen liegt auf *Type* und *Token*-Ebene im juristischen Korpus etwa doppelt so hoch wie im Zeitungstext. Zwar liegt den Zahlen für *–ung-* und *–heit-* / *–keit*-Nominalisierungen eine einfache Suche nach den entsprechende Suffixen zu Grunde. Die Werte in diesen beiden Kategorien decken deshalb auch einen erheblichen Anteil lexikalisierter Sub-

[16] Die Type / Token-Relation variiert stark mit der Textgröße. Die angegebenen Werte sind als Mittelwert über Samples mit je einer Million Textwörtern berechnet. Verzerrungen durch die unterschiedlichen Korpusgrößen sind somit ausgeglichen.

| | | Urteilstexte | | Zeitungstext | |
|---|---|---|---|---|---|
| | *Beispiel* | *Types* | *Textwörter* | *Types* | *Textwörter* |
| **Deverbal** | | | | | |
| Substantivierte Infinitive | *das Betrachten* | 0,69% | 1,38% | 0,37% | 0,54% |
| *-ung*-Derivation | *die Betrachtung* | 10,30% | 4,17% | 4,36% | 1,68% |
| **Deadjektivisch** | | | | | |
| *-heit* / *-keit*-Derivation | *die Richtigkeit* | 1,69% | 0,53% | 0,68% | 0,24% |
| **Gesamtanteil** | | 12,68% | 6.08% | 5.41% | 2.46% |

Tabelle 4.8: Nominalisierungen

stantive mit den entsprechenden Suffixen mit ab, u.a. die im CORTE-Korpus sehr häufigen Substantive *Verwaltung* und *Entscheidung*. Wir gehen allerdings davon aus, dass dieser Effekt im Großen und Ganzen beide Korpora in ähnlichem Maße betrifft, so dass jedenfalls ein Vergleich der Werte in Tabelle 4.8 trotz der Verzerrungen aussagekräftig bleibt (im FR-Korpus gehört beispielsweise *Bundesregierung* zu den häufigsten Substantiven).

Unsere Ergebnisse weisen insgesamt in die gleiche Richtung wie die in Neumann (2009) und Hansen u. a. (2006) diskutierten Beobachtungen. Zwar liegt der Gesamtanteil an Nominalisierungen in unserem Fall etwas niedriger als bei Neumann (2009), die von 7,15% nominalisierten Textwörtern für Urteile bzw. 4,54% im Zeitungstext berichtet. Allerdings berücksichtigt Neumann auch eine größere Zahl derivationaler Nominalisierungsmuster. Dagegen ist das Verhältnis der Nominalisierungshäufigkeiten in unserem Fall noch deutlich unausgewogener als bei Neumann. Ein Grund hierfür könnte darin liegen, dass die von Neumann untersuchten Pressetexte die Gerichtsentscheidungen aus dem Korpus behandeln. Sie sind daher möglicherweise selber in einem stärker juristisch geprägten Duktus verfasst als der im FR-Korpus enthaltene allgemeine Zeitungstext.

**Syntax.** Auf syntaktischer Ebene untermauern praktisch alle von uns untersuchten Indikatoren den qualitativen Eindruck der besonders hohen Komplexität rechtssprachlicher Formulierungen. In Tabelle 4.9 sind auf PreDS-Analysen beruhende Angaben zur Satzlänge, zur durchschnittlichen maximalen Einbettungstiefe, zum Modifikationsgrad und zum Anteil von Passivkonstruktionen in

| | *Urteilstexte* | *Zeitungstext (TIGER-Korpus)* |
|---|---|---|
| **Satzkomplexität** | | |
| *Satzlänge* | 33 (30) | 17 (16) |
| *Prädikate pro Satz* | 1,97 | 1,55 |
| **Einbettungstiefe** | | |
| *Längste Dependenzkette* | 7,18 | 5,45 |
| *Tiefste Satzeinbettung* | 1,66 | 1,35 |
| **Modifikationsgrad** | | |
| *Satz* | 8,44 | 5,48 |
| *Nomen (direkt / transitiv)* | 1,25 / 2,56 | 1,25 / 2,13 |
| *Verb (direkt / transitiv)* | 1,79 / 6,32 | 1,62 / 4,50 |
| **Passivierung** | | |
| *Passivprädikate* | 19,58% | 10,50% |
| *Modales* sein | 4,88% | 0,87% |

Tabelle 4.9: Syntaktische Komplexität

Urteils- und Zeitungstexten aufgeführt. Als Vergleichskorpus diente hier nicht das gesamte FR-Korpus, sondern nur die etwa 40 000 auch ins TIGER-Korpus aufgenommenen Sätze, für die wir PreDS-Strukturen erzeugt haben.

Als erster grober Indikator für syntaktische Komplexität kann die Satzlänge betrachtet werden. Diese liegt im CORTE-Korpus mit durchschnittlich 33 Wörtern etwa beim Doppelten des Wertes für das TIGER-Korpus (17 Wörter). Der Wert für das CORTE-Korpus beruht auf dem Ergebnis automatischer Satzgrenzenerkennung, so dass hier von einer gewissen Zahl extremer Werte aufgrund nicht erkannter Satzgrenzen auszugehen ist. Auch wenn man Satzlängen außer Betracht lässt, die um mehr als zwei Standardabweichungen vom Mittelwert differieren, ergibt sich jedoch mit 30 Wörtern eine nur geringfügig verminderte Durchschnittslänge.

Ein direkteres Maß für die Satzkomplexität stellt die Anzahl der Prädikate in einem Satz dar (jedes Prädikat entspricht einem eigenen Haupt- oder Nebensatz). Auch diese liegt mit 1,97 im CORTE-Korpus deutlich höher als im TIGER-Korpus (1,55). Die im CORTE-Korpus erhöhte durchschnittliche größte Einbettungstiefe (1,66 gegenüber 1,35 im Zeitungstext) liefert einen Anhaltspunkt für die Tendenz zu "Schachtelsätzen", bei denen Nebensätze ihrerseits weitere Nebensätze enthalten. Die Länge der längsten Dependenzkette eines Satzes misst neben der Satzeinbettung auch die Syntax von Konstituenten unter Satzebene. Sie erfasst somit beispielsweise auch die hohe Komplexität von Nominalphrasen, die mit einem Nominalstil einhergeht. Auch diese Größe liegt im CORTE-Korpus mit durchschnittlich 7,18 deutlich höher als im TIGER-Korpus (5,45).

Der Modifikationsgrad (durchschnittliche Anzahl direkter Modifikatoren pro Konstituentenkopf) ist dagegen für Nominalphrasen in beiden Korpora gleich groß. Für Verbalphrasen liegt er im juristischen Korpus wiederum deutlich höher als im TIGER-Korpus. Dies korreliert mit der höheren Zahl von Adverbien und Partikeln in diesem Korpus und dürfte ebenfalls auf die argumentative Funktion vieler Sätze in Urteilsbegründungen zurückzuführen sein.

Unsere Ergebnisse stehen somit im Widerspruch zu der von Wagner diagnostizierten überraschenden Einfachheit der Konstruktionen in ihren Korpora. Die von Neumann und Hansen u. a. für ihr Korpus emittelten Kennzahlen sind zwar aufgrund der verwendeten Maße größtenteils nicht direkt mit den Angaben in Tabelle 4.9 vergleichbar. Der Vergleich mit ihrem Zeitungstext-Korpus ergibt jedoch mit unseren Beobachtungen weitgehend deckungsgleiche Tendenzen. Interessanterweise liegt die in Hansen u. a. (2006) angegebene durchschnittliche Satzlänge von 24,33 Tokens für die Vefassungsgerichtsurteile, der einzige direkt mit Tabelle 4.9 vergleichbare Wert, deutlich unter der für das CORTE-Korpus ermittelten (30 bzw. 33 Tokens). Aufgrund der Größenordnung scheint es durchaus plausibel, dass stilistische Besonderheiten der Sprache des Bundesverfassungsgerichts zumindest eine Ursache für diesen Unterschied darstellen (neben technischen Gründen wie Tokenisierungsdifferenzen oder Problemen bei der Satzgrenzenerkennung, die unser Ergebnis möglicherweise auch nach dem Aussortieren von Ausreißern noch verzerren).

## 4.4 Zusammenfassung und Diskussion

Das Genre *Urteilstext* stellt spezifische Anforderungen an die automatische Verarbeitung, die beim Aufbau eines computerlinguistisch unterstützten juristischen Informationssystems gelöst werden müssen. Wir haben in diesem Kapitel zunächst die Gesamtarchitektur des Definitionsextraktionssystems vorgestellt,

dessen Umsetzung und Auswertung den sprachtechnologischen Schwerpunkt dieser Arbeit darstellt (4.1). Die linguistische Verarbeitung der Eingabetexte bildet in diesem System einen Vorverarbeitungsschritt für die eigentliche Definitionsextraktion.

Die angesprochenen Besonderheiten der behandelten Textsorte kommen hauptsächlich in diesem Schritt zum Tragen. Zunächst wird die Dokumentstruktur ermittelt. Sie ist bei Urteilstexten weitestgehend durch die normativ vorgeschriebene inhaltliche Gliederung bestimmt. Dann wird eine Normalisierung verschiedener – ebenfalls größtenteils genrespezifischer – Textmerkmale (wie Nummerierungen) vorgenommen (4.2.1). Die anschließende, im engeren Sinne linguistische Verarbeitungskette erfolgt nach einer gemeinsamen Satz- und Wortgrenzenerkennung auf einem flachen und einem tiefen Pfad. Während auf dem flachen Pfad elementare lexiko-syntaktische Information (Lemmata und Wortarten, vgl. 4.2.2) angereichert wird, werden auf dem tiefen Pfad syntakto-semantische Dependenzstrukturen erzeugt (sog. PreDS, *Partially resolved dependency structures*, vgl. 4.2.3). Diese unterstützen, wie wir im nächsten Kapitel noch ausführlich darlegen, eine sehr viel konzisere Informationssuche. Während für die flache Verarbeitung vollständig auf generische Komponenten (den in Schmid (1994) beschriebenen TreeTagger und die kommerziell lizensierte Morphologie-Komponente Gertwol der Firma Lingsoft) zurückgegriffen werden konnte, waren für die tiefe Verarbeitung erhebliche Anpassungen des ursprünglichen PreDS-Parsers aus Braun (2003) und Fliedner (2007) notwendig. Unter anderem wurde eine umfassende juristische Zitaterkennung integriert und die Qualität der Parseergebnisse für verschiedene genretypische Konstruktionen und Formulierungsmuster optimiert. Insgesamt wurde so auf den von Fliedner untersuchten juristischen Beispieltexten eine Erhöhung des f-Score um ca. 2,5 Prozentpunkte (nach der von Fliedner vorgeschlagenen Evaluierungsmethodik auf der Basis grammatischer Relations-Tupel) erreicht.

Für die in die in dieser Arbeit dokumentierten Forschungen hatten wir Zugriff auf einen umfangreichen Ausschnitt der von der Firma *juris* geführten Sammlung deutschsprachiger Gerichtsentscheidungen. Für ein Korpus von über 33 000 Entscheidungstexten mit insgesamt mehr als 70 Millionen Textwörtern (der Aufbau des Korpus ist in 4.3.1 genauer beschrieben) liegen PreDS-Strukturen vor. Wir haben dieses Korpus im Rahmen unserer Arbeiten – wie in den nächsten Kapiteln erläutert – vor allem als Referenzressource bei der Auswertung und Optimierung unseres Definitionsextraktionssystems verwendet.

Das geparste Korpus bietet jedoch gleichzeitig eine bisher in diesem Umfang nicht verfügbare Datengrundlage für die empirische Untersuchung von Stilmerkmalen der deutschen Urteilssprache. Wir haben auf seiner Basis verschiedene lexikalische und syntaktische Komplexitätsmerkmale erfasst, die in 4.3.2 erläutert und mit Werten für ein Zeitungstextkorpus sowie den Ergebnis-

sen einiger früherer (auf erheblich kleineren Datenbeständen beruhender) Untersuchungen verglichen sind.

# Kapitel 5

# Automatische Definitionsextraktion aus Urteilstexten

In diesem Kapitel wenden wir uns der Beschreibung der eigentlichen Definitionsextraktion im CORTE-System zu. Wir betrachten diese Aufgabe hier zunächst als eigenständige Zielsetzung und sehen von der Frage nach möglichen Anwendungsszenarios ab, auf die wir dann in Kap. 7 zurückkommen werden. Ein naheliegender Anwendungskontext ist die Ergänzung von Rechtsprechungsdatenbanken um die Möglichkeit einer Suche nach Definitionen für einen Suchbegriff. Denkbar sind aber zum Beispiel auch die Verwendung zur Unterstützung bei der Verschlagwortung von Entscheidungstexten oder – als deutlich weitergehende Nutzungsmöglichkeiten – ein Einsatz im Rahmen der Terminologieextraktion oder Ontologieerstellung.

Zunächst beschreiben wir in 5.1 und 5.2 die Verfahrensschritte bei der Definitionsextraktion im CORTE-System und ihre technische Umsetzung. In 5.3 evaluieren wir dann die Qualität der Extraktionsergebnisse aus unserem vollständig annotierten Datenbestand (siehe Kap. 2) sowie aus dem CORTE-Großkorpus (siehe Kap. 4). Abschließend nehmen wir in 5.4 eine detaillierte Fehleranalyse vor, bei der wir individuelle *false positives* und *false negatives* der Definitionsextraktion auf ihre Ursachen hin untersuchen. Dies ermöglicht uns Rückschlüsse auf Verbesserungsmöglichkeiten und sinnvolle Ergänzungen unseres Suchverfahrens.

Die Betrachtung des Stands der Forschung in Kap. 3 hat gezeigt, dass drei wichtige Parameter in Systemen zur Definitionsextraktion (und zum textbasierten Informationszugriff im allgemeinen) durch folgende Fragen charakterisiert werden können:

1. Auf welche linguistische Information (Analyseebene, Art der Strukturen usw.) wird zurückgegriffen, um Definitionen zu identifizieren?

2. Wird das zur Identifizierung der Definitionen benötigte Wissen (Definitionsmuster o.ä.) von Experten handkodiert (*knowledge engineering*-Ansatz) oder automatisch erworben?

3. Werden – neben den fast immer eingesetzten Mustern zur Beschreibung der gesuchten Formulierungen – noch weitere Wissensquellen für die Definitionsidentifikation genutzt und wie geschieht dies (z.B durch ein Ranking oder die Filterung von Ergebnissen)?

Die Definitionssuche im CORTE-System beinhaltet zwei verschiedene, parallele Schritte, die an den flachen und tiefen Vorverarbeitungspfad (4.2) anknüpfen. Sie nutzen also zwei unterschiedliche Typen linguistischer Information, nämlich Lemmasequenzen bzw. Dependenzanalysen (in Form von PreDS-Strukturen). Wir vergleichen in diesem Kapitel die Ergebnisse beider Schritte und können so den Effekt der linguistischen Analysetiefe auf die Definitionssuche ermitteln.

Sowohl die (lemma)sequenzbasierte als auch die dependenzbasierte Suche stützen sich dabei in der Form, die wir in diesem Kapitel vorstellen, auf rein manuell kodierte Extraktionsregeln. Diese reflektieren die in Kap. 2 korpusbasiert identifizierten definitorischen Formulierungsvarianten. Unsere Evaluation erlaubt es uns somit auch, Stärken und Schwächen des *knowledge engineering*-Ansatzes bei der Definitionsextraktion einzuschätzen.

In Kap. 6 befassen wir uns dann mit Möglichkeiten zur Ergebnisoptimierung durch die Nutzung verschiedener weiterer Wissensquellen (u.a. zur Filterung und Erzeugung eines Rankings) sowie mit der automatisierten Erzeugung weiterer Suchmuster durch ein Bootstrapping-Verfahren.

## 5.1 Sequenzbasierte Extraktion

Viele Ansätze im Bereich der Informations- und Ontologieextraktion nutzen als Kernkomponente der Suche Muster auf der Grundlage von Sequenzen oberflächennaher sprachlicher Elemente (Wortformen, Stämme, Lemmata, Chunks, unter Umständen ergänzt durch morphosyntaktische Zusatzangaben). Dies gilt auch für viele der in 3.2 diskutierten Systeme zur Identifikation von Definitionen. Zu den Vorteilen einer solchen Vorgehensweise gehört es, dass Suchmuster auf einfache Weise direkt aus Korpusbeispielen abgeleitet werden können. Zudem ist nur eine relativ flache (und damit mit wenig Aufwand und zuverlässig durchführbare) linguistische Vorverarbeitung der Textbasis nötig, und der Suchprozess kann mit einfachen Mitteln (z.B. Substringsuche, Positionsindizes oder reguläre Ausdrücke) umgesetzt werden. Andererseits sind sowohl der Trennschärfe als auch den Generalisierungsmöglichkeiten bei der Beschreibung von Formulierungsvarianten durch solche Muster enge Grenzen gesetzt. Wir erläutern im Folgenden die Umsetzung der in Kap. 2 erarbeiteten Definitionsvarianten in sequenzbasierte Definitionsmuster auf lexikalischer Ebene.

### 5.1.1 Spezifikation sequenzbasierter Definitionsmuster

Fixe Wortform- bzw. Stammsequenzen – wie sie von vielen der in 3.2 vorgestellten Ansätze zur Definitionserkennung genutzt werden – sind für die Verarbeitung deutschsprachiger Texte aufgrund der Variabilität der Oberflächenform generell erheblich schlechter geeignet als für das Englische. Probleme entstehen z.B. aufgrund der relativ variablen Wortstellung – etwa der Inversion in Nebensätzen und der Stellungsvarianten im Verbalkomplex – sowie der flexionsbedingten Veränderlichkeit von Lexemen, die in vielen Fällen auch mit Stammvarianten einhergeht.

Als Grundlage für die sequenzbasierte Definitionsextraktion nutzen wir daher sämtliche Annotationen, die im Rahmen der flachen Korpusanalyse (vgl. 4.2.2) erzeugt wurden. Neben der Zuordnung zu einem Dokument und einem Satz sind für jedes Token somit folgende Informationsebenen verfügbar:

- String
- Lemma
- Part of Speech
- Position im Satz

Unsere Suchmuster lassen Angaben zu allen vier Ebenen zu. Insgesamt muss in jedem Muster für mindestens ein Token String, Lemma oder Part of Speech spezifiziert werden.

Für die ersten drei Ebenen werden Strings angegeben, wobei eine Links- oder Rechtstrunkierung zur Markierung von Suffix bzw. Präfixzeichenfolgen möglich ist. Auf String- und Lemma-Ebene ermöglicht dies u.a. eine Generalisierung über die alternativen Formen von Pronominaladverbien für Nahes und Fernes (z.B. *hierunter* vs. *darunter*) und die verschiedenen Flexionsformen mit dem bestimmten Artikel verschmolzener Präpositionen (z.B. *zur* vs. *zum*). Auf Part of Speech-Ebene nutzen wir die Möglichkeit der Trunkierung für die einheitliche Beschreibung der attribuierenden und substituierenden Formen des Relativpronomens, die in dem von uns verwendeten STTS[1], unterschiedlich getaggt werden (vgl. das Muster in Abb. 5.1).

Auf Positionsebene können Einschränkungen über Tokenpaare formuliert werden, die eine direkte Abfolge oder eine bloße Präzedenz fordern. Indem solche Einschränkungen nur für ausgewählte Paare festgelegt werden, kann Stellungsvarianten Rechnung getragen werden. Zur kompakten Beschreibung einfacher Varianten sind zudem auf allen drei Ebenen Disjunktionen von Bedingungen zulässig.

[1] **S**tuttgart-**T**übingen **T**agset für das Deutsche, beschrieben in Schiller u. a. (1999)

```
tokens       =>  #unter#APPR|#%unter#PAV,#zu#PTKZU,
                 #verstehen#VVINF,#sein,##PREL%
ordering     =>  p1<p2,p2<p3!
description  =>  unter zu verstehen sein, relpron
```

Abbildung 5.1: Sequenzbasiertes Suchmuster für Definitionen auf der Basis des Prädikats *verstehen unter*

Abb. 5.1 zeigt ein sequenzbasiertes Suchmuster zur Identifikation von Definitionen, die mit dem Prädikat *verstehen unter* (in der modalen Passivumschreibung *zu verstehen sein*) und einem modifizierenden Relativsatz zur Angabe einer *differentia* getroffen werden. Es sind insgesamt fünf Token spezifiziert:

1. die Präposition *unter* zur Einleitung des Definiendum, alternativ als echte Präposition (APPR) oder (linkstrunkiert durch ein %-Zeichen) als Bestandteil eines Pronominaladverbs (PAV, *da-* bzw. *hierunter*),
2. die Infinitivpartikel *zu*,
3. das Vollverb *verstehen*,
4. das Auxiliar *sein* und
5. ein beliebiges Relativpronomen (ohne Angabe einer Wortform oder eines Lemma) als notwendiger Bestandteil des Definiens.

Zur Anordnung der Token wird zusätzlich festgeschrieben, dass die Präposition bzw. das entsprechende Pronominaladverb (und somit das Definiendum) der Infinitivpartikel vorausgeht und diese wiederum unmittelbar vor dem Vollverb steht. Solche Wortform / Lemma / POS-Angaben mit Anordnungsbeschränkungen lassen ein gewisses Maß an Abstraktion von der sprachlichen Oberflächenform zu. So kann durch das Muster in Abb. 5.1 die intendierte Definitionsform z.B. in unterschiedlichen Wortstellungen sowie in infiniter und verschiedenen finiten Formen erkannt werden.

Die gewählte Beschreibungsebene trägt den Gegebenheiten der deutschen Sprache erheblich besser Rechnung als reine Wortform- oder Stammsequenzen

und erlaubt somit konzisere und generellere Suchmusterspezifikationen. Andererseits bleibt die Ausdrucksmächtigkeit von Suchmustern auch auf der Basis der beschriebenen linguistischen Informationstypen noch stark eingeschränkt. Beispielsweise lassen sich Aktiv- und Passivvarianten desselben Prädikats nur durch separate Muster beschreiben, und es können keine Einschränkungen zum Ausdruck gebracht werden, die sich auf phrasale Einheiten oder Abhängigkeiten zwischen Wörtern bzw. Wortgruppen beziehen. Dies bedeutet insbesondere, dass die funktionalen Bestandteile prädikatbasierter Definitionen nicht durch sequenzbasierte Muster identifiziert werden können, da es sich bei diesen um (meist komplexe) Phrasen handelt, die nicht durch ihre Position oder bestimmte Signalwörter, sondern durch ihre grammatische Funktion bezüglich des Definitionsprädikats charakterisiert sind (vgl. 2.3.2).

### 5.1.2 Suche

Um Definitionsmuster wie das in Abb. 5.1 wiedergegebene zur Identifikation von Definitionen in einem Textbestand zu verwenden, ist es erforderlich, dass dieser in durchsuchbarer Form abgelegt wird und die Suchmuster in ausführbare Anfragen einer passenden Anfragesprache überführt werden.

Für die Speicherung von Textdaten mit mehrschichtigen Annotationen und die Suche in solchen Beständen existieren verschiedene, in unterschiedlichen Forschungs- und Anwendungszusammenhängen entwickelte Lösungen. So bietet etwa die am Institut für maschinelle Sprachverarbeitung der Universität Stuttgart entwickelte *Corpus Workbench* (Christ (1994)) die Möglichkeit, für Anfragen auf der Basis von Korpuspositionen mit beliebigen Attributen *keyword in context*-Listen zu erzeugen. Sie ist stark auf die Bedürfnisse korpuslinguistischer und lexikographischer Untersuchungen ausgerichtet (Ähnliches gilt für die *Sketch Engine* der Firma *Lexical Computing Ltd.*, Kilgarriff u. a. (2004)). Dagegen sind etwa die *Apache Lucene*-Bibliothek (Gospodneti und Hatcher (2004)) und die Speicher- und Suchfunktionalität der GATE-Plattform (Cunningham (2000)) stärker an den Aufgabenstellungen des Informationszugriffs orientiert. Sie bieten z.B. spezialisierte Indizierungsmöglichkeiten und erlauben eine Einbindung in Gesamtsysteme über Programmierschnittstellen (APIs).

Im Rahmen der hier beschriebenen Studie standen jedoch weder lexikographische Fragestellungen noch Effizienz- und Architekturgesichtspunkte im Vordergrund. Für den beabsichtigten Schwerpunkt auf der Evaluation unterschiedlicher Suchmuster unter verschiedenen Gesichtspunkten erschienen uns dagegen ein in hohem Maße transparenter und kontrollierbarer Suchprozess, die Möglichkeit zur flexiblen Integration von Suchergebnissen mit Metadaten und zusätzlichen Annotationen sowie möglichst freie Abfragemöglichkeiten als be-

sonders relevante Faktoren. Wir haben uns daher entschieden, die sequenzbasierte Definitionsextraktion nicht mittels eines der erwähnten spezialisierten Systeme, sondern vollständig innerhalb der bereits in Kap. 4 erwähnten relationalen MySQL-Datenbank umzusetzen. Suchmuster auf der Grundlage der oben diskutierten Information und Kombinationsmöglichkeiten lassen sich direkt und automatisiert in SQL-Anfragen umschreiben. Abb. 5.2 zeigt die aus dem Muster in Abb. 5.1 erzeugte Anfrage.

```
SELECT DISTINCT t1.sent_id
 FROM
  tokens t1 JOIN tokens t2 ON (t1.sent_id=t2.sent_id)
  JOIN tokens t3 ON (t1.sent_id=t3.sent_id)
  JOIN tokens t4 ON (t1.sent_id=t4.sent_id)
  JOIN tokens t5 ON (t1.sent_id=t5.sent_id)
 WHERE
  (((t1.lemma='unter' AND t1.pos='APPR') OR        # (1)
    (t1.lemma LIKE '%unter' AND t1.pos='PAV')) AND
   (t2.lemma='zu' AND t2.pos='PTKZU') AND
   (t3.lemma='verstehen' AND t3.pos='VVINF') AND
   (t4.lemma='sein') AND
   (t5.pos LIKE 'PREL%'))                          # (2)
 AND
  ((t1.sentenceposition<t2.sentenceposition) AND   # (3)
   (t3.sentenceposition=(t2.sentenceposition+1)))
 AND                                               # (4)
  (NOT(t1.id=t4.id)  AND  NOT(t2.id=t5.id)  AND
   NOT(t3.id=t4.id)  AND  NOT(t2.id=t4.id)  AND
   NOT(t1.id=t3.id)  AND  NOT(t2.id=t3.id)  AND
   NOT(t4.id=t5.id)  AND  NOT(t1.id=t2.id)  AND
   NOT(t1.id=t5.id)  AND  NOT(t3.id=t5.id))
```

Abbildung 5.2: SQL-Anfrage für das Suchmuster in Abb. 5.1 (die Anfrage bezieht sich auf eine Tabelle, deren Datensätze die Tokens im Korpus, jeweils mit den Attributen String, Lemma, Part of Speech (`pos`), Satzzugehörigkeit (`sent_id`) und Position im Satz (`sentenceposition`) repräsentieren.)

Konjunktion und Disjunktion werden durch die SQL-Operatoren AND und OR ausgedrückt (1), den Abgleich mit trunkierten Zeichenfolgen leistet der LIKE-Operator (2), die Anordnungsbeschränkungen werden als Positionsvergleiche über die einzelnen Token ausgedrückt (3). Durch eine Zusatzbedingung (4) wird sichergestellt, dass jedes in der Anfrage verwendete Alias auf einen an-

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **String** | Darunter | ist | eine | Lage | zu | verstehen | , |
| **Lemma** | dar**unter** | **sein** | ein | Lage | **zu** | **verstehen** | , |
| **POS** | **PAV** | VAFIN | ART | NN | **PTKZU** | **VVINF** | $, |
| **Position** | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

| | | | | | |
|---|---|---|---|---|---|
| **String** | in | der | … | politische | Verfolgungsmaßnahmen |
| **Lemma** | in | d | | politisch | Verfolgungsmaßnahme |
| **POS** | APPR | **PREL** | | ADJA | NN |
| **Position** | 8 | 9 | … | 17 | 18 |

| | | | | | |
|---|---|---|---|---|---|
| **String** | … | nicht | auszuschließen | sind | . |
| **Lemma** | | nicht | ausschließen | sein | . |
| **POS** | | PTKNEG | VVIZU | VAFIN | $. |
| **Position** | … | 37 | 40 | 41 | 42 |

Abbildung 5.3: Ableitung des Suchmusters in Abb. 5.1 aus Bsp. (5.1)

deren Datensatz Bezug nimmt, die einzelnen im Suchmuster charakterisierten Token also tatsächlich unterschiedlich sind.

### 5.1.3 Genutzte Muster

Die Umsetzung der in Kap. 2 identifizierten definitorischen Formulierungen in sequenzbasierte Definitionsmuster kann auf weite Strecken systematisch durch Tilgung von Material aus den analysierten Korpusbeispielen erfolgen. So basiert beispielsweise das in Abb. 5.1 wiedergegebene Muster auf (5.1).

(5.1) Darunter ist eine Lage zu verstehen, in der dem Ausländer vor seiner Ausreise im Heimatstaat politische Verfolgungsmaßnahmen zwar – noch – nicht mit beachtlicher Wahrscheinlichkeit drohen, nach den gesamten Umständen jedoch auf absehbare Zeit nicht hinreichend sicher auszuschließen sind, weil Anhaltspunkte vorliegen, die ihren Eintritt als nicht ganz entfernt erscheinen lassen.

(Sächsisches Oberverwaltungsgericht 4. Senat, 28. August 2001, AZ A 4 B 4388/99, juris)

Aus diesem Satz kann es durch Übernahme der in Abb. 5.3 hervorgehobenen Elemente partiell abgeleitet werden.

Ergänzende Einschränkungen sowie Verallgemeinerungen erfordern die Betrachtung des grammatisch zulässigen Variationspotentials sowie weiterer Korpusbeispiele. So wurden für das Muster in Abb. 5.1 zwar die relative Position des Präpositionaladverbs zum Infinitiv sowie die Abfolge *zu+Infinitiv* als Anordnungsbeschränkungen übernommen. Die Stellung des Auxiliarverbs *sein* relativ zu den anderen Elementen wurde dagegen offen gelassen, weil dieses in Nebensätzen (anders als in (5.1)) dem Hauptverb nachfolgt. Auf der Basis anderer Korpusbeispiele wurde außerdem als alternative Realisierungsmöglichkeit anstelle des Präpositionaladverbs *da-/hierunter* die attributive Präposition *unter* zugelassen.

Entscheidungen über Zusatzbedingungen und Generalisierungen wie die angesprochenen sind durch Beispiele und Sprachkompetenz jedoch in der Praxis nicht vollständig determiniert (so hätte beispielsweise in Abb. 5.1 die unpersönliche Passivumschreibung mit *man* berücksichtigt werden können oder die stark präferierte Abfolge *unter* < *Relativpronomen* zusätzlich festgeschrieben werden können). Sie sind daher ohne Richtlinien bei der Spezifikation einer größeren Zahl von Suchmustern kaum einheitlich zu treffen. Wir haben bei der Zusammenstellung der hier verwendeten sequenzbasierten Suchmuster Anordnungsbeschränkungen nur dann übernommen, wenn ohne sie die Grammatikalität der definitorischen Formulierung zweifelhaft erschien. Zudem haben wir, neben der illustrierten Alternative *Pronominaladverb / attributive Präposition*, grundsätzlich Verallgemeinerungen auf der Basis folgender Alternationen getroffen, auch wenn diese in unseren Entwicklungsdaten nicht attestiert waren:

- *dass* vs. *ob* zur Einleitung indirekter Rede, soweit die Alternative semantisch offen stand,
- *werden*-Passiv / modale *sein*-Umschreibung,
- getrennte / ungetrennte Variante trennbarer Präfixverben (mit Anordungsbeschränkung im Fall *getrennt*).

Während der erste Fall durch disjunktive Angaben innerhalb eines Musters repräsentiert werden kann, erfordern die beiden anderen Alternationen jeweils die Verdoppelung eines Suchmusters. Bei der Modellierung kopulabasierter Definitionen haben wir zudem systematisch unterschiedliche Realisierungen des Prädikativum (Nomen sowie pronominale Varianten, etwa *derjenige* und *alle*) und des Modifikators (u.a. als Relativsatz, Genitivattribut, Konditionalsatz oder Modalsatz) sowie verschiedene zusätzliche Signalwörter (etwa *insbesondere* und *unter anderem*) berücksichtigt, wozu ebenfalls jeweils die Wiederholung ansonsten identischer Muster notwendig war.

| **Prädikat** | **Konstruktionen** | **Muster** |
|---|---|---|
| *sein* | 2 | 21 |
| *verstehen* | 3 | 6 |
| *ansehen* | 2 | 4 |
| *ausschließen* | 2 | 4 |
| *voraussetzen* | 2 | 4 |
| *vorliegen* | 2 | 4 |
| *gelten* | 2 | 3 |
| *ausreichend sein* | 2 | 2 |
| *betrachten* | 2 | 2 |
| *gehören (zu)* | 2 | 2 |
| *werden* | 2 | 2 |
| *anerkennen* | 1 | 2 |
| *annehmen* | 1 | 2 |
| *ausgehen (von)* | 1 | 2 |
| *darstellen* | 1 | 2 |
| *müssen* | 1 | 2 |
| *zukommen* | 1 | 2 |
| *bedeuten, sich begnügen müssen, besagen, der Fall sein, einbeziehen (in), entfallen, entscheidend sein, erforderlich sein, erfordern, gegeben sein, Gegenstand sein, gekennzeichnet sein, genießen, genügen, genügen + Dat., gleichgültig sein, haben, sich handeln (um), maßgebend sein, sich richten (gegen), schützen, sprechen (von), umfassen, umschreiben, verlangen, Voraussetzung sein, zählen (zu)* | je 1 | je 1 |
| Gesamtzahl | 56 | 93 |

Tabelle 5.1: Zuordnung sequenzbasierter Muster zu Definitionsprädikaten

Tabelle 5.1 führt für die in Kap. 2 (Tabelle 2.4) genannten Definitionsprädikate (mit Ausnahme der in unseren Experimenten nicht berücksichtigten Prädikate zur Angabe erläuternder / kommentierender Zusatzinformation) jeweils auf, in wie vielen verschiedenen Konstruktionen zur Angabe von Definiendum und Definiens diese auftreten und wie viele sequenzbasierte Muster wir zu ihrer Modellierung verwenden. Insgesamt ergibt sich aus den genannten Gründen

eine verglichen mit der Zahl der in Kap. 2 diskutierten Formulierungstypen erheblich höhere Anzahl sequenzbasierter Muster.

## 5.2 Dependenzbasierte Extraktion

Wenn für die Textbasis außer flachen, oberflächennahen linguistischen Analyseergebnissen auch strukturelle Information aus tieferen Analyseschritten (partielle oder vollständige syntaktische Strukturen oder semantische Repräsentationen) zur Verfügung steht, ist eine kompaktere und zugleich genauere Beschreibung von Formulierungstypen möglich als mit sequenzbasierten Mustern. Gruppierungen und Abhängigkeiten zwischen einzelnen Elementen, die in sequenzbasierten Mustern nur indirekt (über Abfolgebeschränkungen) kodiert werden können, lassen sich auf der Basis tieferer Analysen direkt beschreiben. Gleichzeitig können Stellungsvarianten sowie – je nach Normalisierungsgrad der jeweiligen Analyse – auch andere semantisch gleichwertige Formulierungsalternativen (etwa Aktiv-Passiv-Alternationen) durch einheitliche Muster erfasst werden.

Andererseits gestalten sich die Vorverarbeitung der Textbasis, die Verwaltung und Suche in den erzeugten Textrepräsentationen sowie die Erstellung von Suchmustern komplexer als im Fall flacher, einzelwortbasierter Analysen.

Einige der in Kap. 3 vorgestellten Ansätze zur Definitionsextraktion sowie ein großer Teil der jüngeren IE-Systeme nutzen Suchmuster auf der Grundlage von Dependenzanalysen. Wir gehen im Folgenden auf die in unseren Experimenten genutzten dependenzbasierten Definitionsmuster ein. Methodologisch verfolgen wir (zunächst) wie für die sequenzbasierten Muster einen *knowledge engineering approach* und entwickeln daher einen Rahmen, in dem sich die in Kap. 2 erarbeiteten definitionsspezifischen linguistischen Attribute zur effizienten manuellen Spezifikation von Definitionsmustern nutzen lassen. In Kap. 6 diskutieren wir dann ein Bootstrapping-Verfahren zum automatischen Erwerb weiterer Suchmuster.

### 5.2.1 Spezifikation dependenzbasierter Definitionsmuster

Prädikatbasierte Definitionen können im wesentlichen anhand von zwei Dimensionen beschrieben werden (vgl. 2.3):

1. Welches Definitionsprädikat wird verwendet?
2. Wie werden Definitionsbestandteile in der syntaktischen Umgebung des Prädikats realisiert?

Im Hinblick auf die Realisierung der Definitionsbestandteile lassen sich dabei zwei verschiedene Aspekte unterscheiden: Welche syntaktischen Einheiten können in Kombination mit einem gegebenen Prädikat in Definitionen auftauchen, und wie sind diese den funktionalen Definitionsbestandteilen (z.B. Definiendum und Definiens) zugeordnet?[2] Allerdings tragen häufig noch weitere linguistische Zusatzindikatoren zur Unterscheidung zwischen definitorischen und nicht-definitorischen Auftreten derselben Konstruktion bei.

Abb. 5.4 und Abb. 5.5 illustrieren diese Einteilung anhand der PreDS von zwei Definitionsinstanzen mit dem Prädikat *verstehen*, einmal in der modalen Passivumschreibung durch *zu... sein* und einmal in einer echten Passivkonstruktion (vgl. (5.1) bzw. (5.2), s.u.). Hervorgehoben sind die gemeinsamen strukturellen Bestandteile beider Sätze, die sie als Instanzen derselben definitorischen Formulierung ausweisen. Dass die zweite PreDS eine modale Umschreibung repräsentiert, ist in dem Merkmal *modSein* des Definitionsprädikats *verstehen* kodiert.[3] Neben dem Definitionsprädikat sind der syntaktische Rahmen (Tiefenobjekt und modifizierende Präpositionalphrase mit *unter*) und Zuordnung seiner Elemente zu den Definitionsbestandteilen (Definiendum und Definiens) zu erkennen.

Bei dem gelb gedruckten, beiden Instanzen gemeinsamen Attribut *Passiv* des Definitionsprädikats handelt es sich um einen Zusatzindikator für das Vorliegen einer Definition: Durch Depersonalisierung wird ein Anspruch auf Allgemeingültigkeit der Aussage zum Ausdruck gebracht. Aufgrund des relativ hohen linguistischen Abstraktionsgrads in PreDS-basierten Textrepräsentationen führen funktional gleichwertige Alternativen bei solchen Zusatzindikatoren oft nur zu relativ kleinen, lokalen Änderungen der PreDS, die die erstgenannten Hauptbeschreibungsebenen nicht betreffen. Im Beispiel etwa würde sich eine semantisch und funktional äquivalente Formulierung im unpersönlichen Aktiv (*Unter... versteht man...*) nur hinsichtlich des Passivmerkmals und des zusätzlichen Tiefensubjekts *man* von den beiden angeführten Beispielen unterscheiden. Die sonstige Struktur auf PreDS-Ebene – und somit insbesondere die Identifikation des Definiens als Tiefenobjekt – bliebe aufgrund der Aktiv-Passiv-Normalisierung unverändert.

In Abb. 5.6 ist das (u.a. dem sequenzbasierten Muster aus Abb. 5.1 entsprechende) dependenzbasierte Muster zur Erkennung von Definitionen mit dem Prädikat *verstehen* wiedergegeben, mit dem beispielsweise die Definitionen in (5.1) und (5.2) erkannt werden können.

[2] Wie in Kap. 2 erläutert füllen die Definitionsbestandteile typischerweise semantische Argumentstellen des Definitionsprädikats. Ihre syntaktische Realisierung hängt somit vom Linking-Muster des Definitionsprädikats ab.

[3] Siehe zu den Einzelheiten des PreDS-Formats Kap. 4.

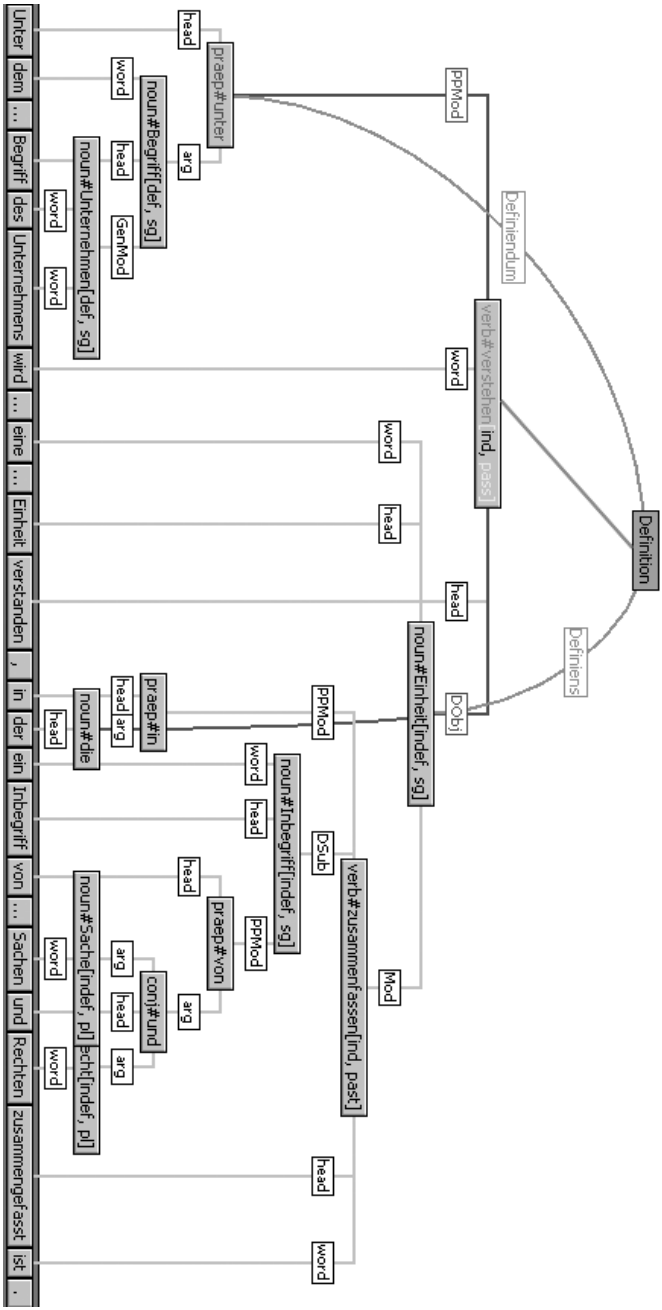



Abbildung 5.4: Dependenzbasiertes Definitionsmuster (1)

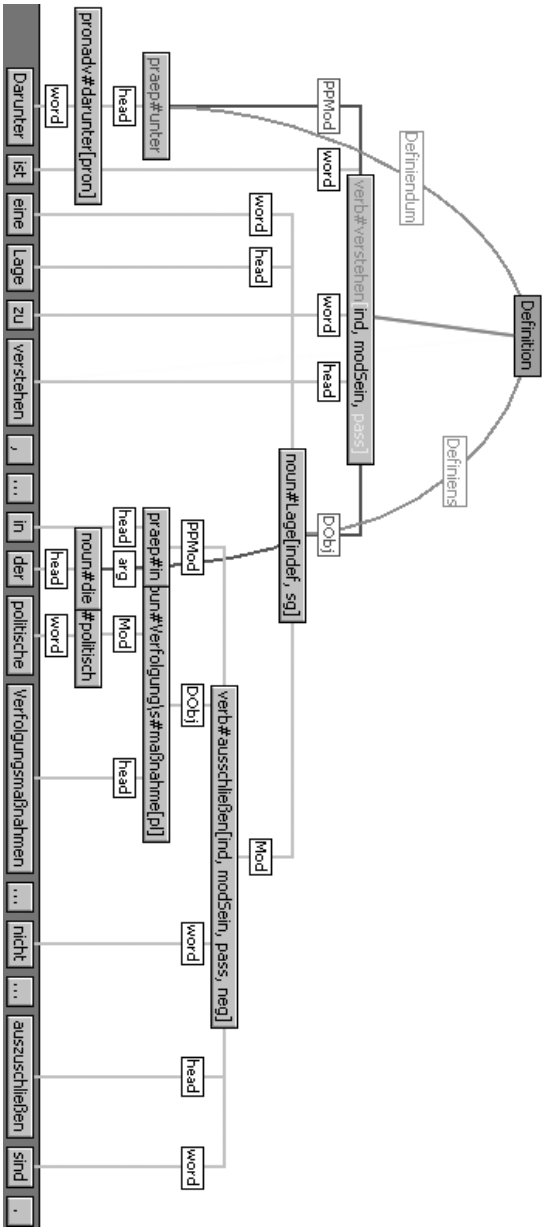



Abbildung 5.5: Dependenzbasiertes Definitionsmuster (2)

```
<pattern>
    <key>verstehen</key>
    <frames>
        <frame id="DObj:Rcl-PP" args="PREP:unter">
            <mapping id="PP:defined_DObj:defining"
                     args="PREP:unter"/>
            <condition>PASS and INDPRES</condition>
        </frame>
        <frame id="DObj:RCl-alsNP">
            <mapping id="alsNP:defined_DObj:defining">
            <condition>PASS and INDPRES</condition>
        </frame>
    </frames>
</pattern>
```

Abbildung 5.6: Dependenzbasiertes Muster für Definitionen mit dem Prädikat *verstehen*

(5.2) Unter dem vermögensrechtlichen Begriff des Unternehmens wird nach ständiger Rechtsprechung des Bundesverwaltungsgerichts in Anlehnung an den Unternehmensbegriff des Handelsrechts eine organisatorische Einheit verstanden, in der ein Inbegriff von gemeinsamen wirtschaftlichen Zwecken dienenden Sachen und Rechten sowie sonstigen wirtschaftlichen Werten, wie unternehmerische Erfahrung, Geschäftsbeziehung, Ruf, Kundenstamm, zusammengefasst ist (vgl. BVerwG, Urteil vom 28. März 2001 - 8 C 6/00 - VIZ 2001, 609).

(VG Gera 5. Kammer, 24. Oktober 2001, AZ 5 K 14/99 GE, juris)

Die Grundstruktur unserer dependenzbasierten Definitionsmuster spiegelt die erläuterten Beschreibungsebenen wider: Es wird jeweils (1) das Definitionsprädikat (*key*), (2) ein syntaktischer Rahmen (*frame*) und (3) dessen Abbildung auf Definitionsbestandteile (*mapping*) angegeben. Zusätzlich können dann (4) weitere Randbedingungen (*conditions*) formuliert werden.

Zur Spezifikation verwenden wir ein XML-basiertes Format. Ein Definitionsprädikat kann mit verschiedenen syntaktischen Rahmen und Zuordnungen von Definitionsbestandteilen auftreten, und derselbe Rahmen kann von verschiedenen Prädikaten instantiiert werden. Zudem können die Elemente eines syntakti-

schen Rahmens in verschiedenen Fällen unterschiedlich auf Definitionsbestandteile abzubilden sein. Unser Suchmusterformat läßt daher für jeden *key* eine beliebige Anzahl von *frame+mapping* Kombinationen zu, die durch symbolische Abkürzungen spezifiziert werden und jeweils mit zusätzlichen *conditions* veknüpft werden können. Die vollen Ausdrücke für *frames* und *mappings* sind in einer eigenen Datei zentral spezifiziert und werden bei der Erzeugung ausführbarer Suchanfragen (s.u.) entsprechend den symbolischen Referenzen importiert.

In dem in Abb. 5.6 angeführten Definitionsmuster sind zu dem Prädikat *verstehen* die Rahmen *unter-PP + Tiefenobjekt mit Relativsatz* (die Präposition wird als eigenes Argument angegeben) und *als-NP + Tiefenobjekt mit Relativsatz* mit entsprechenden *mappings* spezifiziert. Als Zusatzbedingung wird jeweils ein Prädikat im Indikativ Präsens Passiv (einschließlich der Umschreibung durch modales sein) gefordert. Die oben erwähnte (in unseren Entwicklungsdaten allerdings nicht attestierte) Formulierungsvariante *Unter / als . . . versteht man . . .* könnte für beide *frames* durch disjunktive Verknüpfung einer entsprechenden Bedingung mit dem Passiv-Attribut in der jeweiligen *condition* abgedeckt werden.

### 5.2.2 Suche

Für die Suche in tiefer strukturierten syntaktischen Annotationen existieren – verglichen mit der Suche in flach, d.h. wort- oder phrasenweise annotierten Korpora – gegenwärtig relativ wenige umfassende Lösungen. Kein Ansatz hat sich bisher als Standard im Kontext praktischer Anwendungen für den Informationszugriff etabliert.

Allgemeine Informationszugriffskomponenten wie die bereits erwähnte *Lucene*-Bibliothek sind nicht für die Nutzung syntaktisch analysierter Korpora ausgelegt. Die Indizierung und Durchsuchung von Baumstrukturen kann nur auf der Basis einer Abbildung ausgewählter Informationstypen in flache Strukturen erfolgen.

Linguistisch orientierte Suchwerkzeuge für Baumbanken sind dagegen meist auf ein bestimmtes Repräsentationsformat (entweder das in der jeweiligen Baumbank verwendete Analyseformat oder, wie im Falle von *TigerSearch* (Lezius (2002)), ein generisches Importformat) spezialisiert und stellen auf dieses zugeschnittene Anfragefunktionalitäten zur Verfügung. Sie sind in der Regel für die direkte Nutzung durch Linguisten optimiert (im Falle von *TigerSearch* etwa durch eine Schnittstelle zur graphischen Spezifikation von Anfragen).

Die Nutzung solcher Werkzeuge für die PreDS-basierte Definitionssuche hätte also in jedem Fall die Implementierung von Importfiltern und spezialisierten Anfrageschnittstellen erforderlich gemacht. Bei der Nutzung von (beispielswei-

se) *Lucene* wäre zudem von Anfang an eine Einschränkung der für die Suche zugreifbaren linguistischen Information zu treffen gewesen, während bei der Verwendung von Baumbank-Werkzeugen Schwierigkeiten bei der Integration mit anderen Systemkomponenten zu erwarten waren.

Im Rahmen der hier diskutierten Untersuchung bestanden die Hauptanforderungen an den Suchprozess jedoch in einem möglichst flexiblen Zugang zu allen in der linguistischen Analysephase ermittelten Informationen, möglichst großer Transparenz und Kontrollierbarkeit sowie problemloser Interoperabilität aller experimentellen Systemkomponenten. Bei der Umsetzung der Suche haben wir uns daher entschieden, das auch im PreDS-Parser genutzte XML-basierte Datenformat durchgängig beizubehalten und den Musterabgleich auf der Basis von XPath-Ausdrücken (also unter Nutzung von W3C-Standardtechnologie, vgl. Clark und DeRose (1999)) zu realisieren. Hierzu war zum einen eine Übersetzung der Definitionsmuster in XPath-Ausdrücke erforderlich, zum anderen wurde eine Lösung zur XML-basierten Datenhaltung mit Suchfunktionalität benötigt.

#### Übersetzung von Definitionsmustern

Die Übersetzung von Definitionsmustern in XPath-Anfragen erfolgt in zwei Schritten. Zunächst werden für jedes `pattern`-Element (vgl. Abb. 5.6) die verwendeten *frames* und *mappings* importiert, instantiiert und zu einer linguistisch motivierten Zwischenrepräsentation zusammengefügt. Aus dieser werden dann XPath-Ausdrücke erzeugt.

Abb. 5.7 zeigt die Ausdrücke die mit den Bezeichnern `PP:defined_-DObj:defining` und `DObj:Rcl-PP` verknüpft sind (1), dann die aus diesen und dem Muster in Abb. 5.6 erzeugte Zwischenrepräsentation (2) sowie die letztendlich generierten XPath-Ausdrücke (3).

Für diese mehrstufige Vorgehensweise sprechen zwei wichtige Gründe: Zum einen kann durch die Verwendung entsprechender Abstraktionen auf der Ebene der Zwischenrepräsentationen die Modellierung von Definitionsformulierungen zunächst rein linguistisch motiviert und unabhängig von Implementierungsgesichtspunkten erfolgen. Die Suchmuster sind somit weitgehend theorieneutral und würden bei Verfügbarkeit z.B. eine relativ unproblematische Umstellung auf die Nutzung vom PreDS-Format abweichender Analysestrukturen ermöglichen.

Zum anderen kann die Erzeugung von XPath-Ausdrücken modular gegen eine Übersetzung in Anfragen auf der Basis beliebiger anderer (z.B. nicht XML-basierter) Datenformate ausgetauscht werden. So können einmal spezifizierte Definitionsmuster mit unterschiedlicher Suchtechnologie genutzt werden (wir

(1)

```
<frame id="DObj:Rcl-PP">
    <description>KEY + DObj:Rcl-PREPPP</description>
    <query>
        PREFIX[@key="KEY" and PPMODSTEM%PREP and RELRCREL%DObj]]
    </query>
</frame>

<mapping id="PP:defined_DObj:defining">
    <item field="defined" keys="span,stem,context">
        PPMODSTEM%PREP
    </item>
    <item field="defining" keys="span,stem,context">DOBJ</item>
</mapping>
```

$\longrightarrow$

(2)

```
query: @key="verstehen" and DOBJ and PPMODKEY%unter
    and PASS and INDPRES
defined: PPMODKEY%unter
defining: DOBJ
```

$\longrightarrow$

(3)

```
query: word[@key="verstehen" and PPMod/word[@key="unter"]
      and DObj and attrs/pass and attrs/ind and not(attrs/past)]
defined: PPMod/word[@key="unter"]/arg/word
defining: DObj/word
```

Abbildung 5.7: *Frame* und *mapping* für die erste in Abb. 5.6 spezifizierte Konstellation

gehen im Folgenden kurz auf die Verwendung eines relationalen Datenmodells mit SQL-basierter Suche ein).

Dabei entscheidet allerdings die Expressivität der eingesetzten Anfragesprache darüber, welche Ausdrucksmöglichkeiten für die konkrete Umsetzung der in den Zwischenrepräsentationen verwendeten linguistischen Abstraktionen zur Verfügung stehen. XPath bietet hier sehr weitreichende Möglichkeiten: XPath-Ausdrücke beschreiben Mengen von XML-Dokumentbäumen. Sie er-

lauben dabei eine flexible Adressierung von Knoten durch Pfadausdrücke mit Aufwärts- und Abwärtsschritten sowie Wildcards für einzelne Schritte oder beliebige Schrittsequenzen. Pfadausdrücke können außerdem Prädikate über Knoten enthalten. Diese können neben einem reichen Inventar von Funktionen wiederum rekursiv volle XPath-Ausdrücke in konjunktiver und disjunktiver Verknüpfung sowie Negationen enthalten.

Bei der Übersetzung der von uns genutzten Definitionsmuster wird von diesen Möglichkeiten jedoch nur eingeschränkt Gebrauch gemacht. Die Muster beschreiben den Prädikatknoten und einen relativ lokalen syntaktischen Kontext, und auch die verwendeten abstrakten linguistischen Prädikate beziehen sich meist nur auf Eigenschaften eines einzelnen PreDS-Knotens und seiner unmittelbaren Umgebung. Diese können allerdings logisch komplex sein, so dass die Möglichkeiten zur Verknüpfung von XPath-Prädikaten für die Umsetzung der Definitionsmuster eine erheblich größere Rolle spielen als die flexible Adressierung von Knoten in XML-Dokumentbäumen.

#### (a) Datenhaltung und Suche

Sowohl kommerzielle Anbieter als auch verschiedene Open Source-Projekte arbeiten seit geraumer Zeit an Lösungen für die datenbankbasierte Verwaltung von XML-Daten mit Abfragemöglichkeiten auf der Grundlage von XPath (bzw. der auf XPath basierenden Abfragesprache XQuery, Chamberlin und Robie (2008)). Dabei werden unterschiedliche Ansätze verfolgt. So integrieren beispielsweise Oracle und MySQL XML-Module in ihre relationalen Datenbank-Management-Systeme (sog. *XML-enabled DBMS*), während etwa die Apache Foundation, die Software AG sowie das (inzwischen von Oracle aufgekaufte) Unternehmen Sleepycat Lösungen entwickeln, deren interne Speicherung und Verarbeitung vollständig auf das XML-Datenmodell gestützt und auf Baumstrukturen optimiert ist (sog. *native XML DBMS*).

Zu Beginn unserer Arbeiten hatte sich jedoch keines der entwickelten Systeme am Markt oder in größeren Forschungsprojekten etabliert.[4] Die verfügbaren *XML-enabled DBMS* boten zudem nur eine eingeschränkte Unterstützung der XPath und XQuery-Standards. Experimente mit den frei verfügbaren nativen Lösungen BerkeleyDBXML (*Sleepycat*) und eXist ergaben dagegen Stabilitätsprobleme bei der Verwaltung großer Datenmengen sowie eine für unsere Zwecke zu geringe Transparenz und Kontrollierbarkeit des Datenzugriffs (z.B. fehlende Rückmeldung und Interaktionsmöglichkeiten bei lang laufenden Anfragen).

[4] Unseres Wissens haben sich an dieser Situation auch zum jetzigen Zeitpunkt (Sommer 2009) noch keine wesentlichen Änderungen ergeben.

Wir nutzen für die Ausführung von XPath-Anfragen daher keine datenbankgestützte Lösung, sondern verwenden eine dateibasierte Architektur, dynamisch generierte XSLT-Skripte sowie die im CPAN-Archiv verfügbaren Bibliotheken XML::LibXSLT (Ausführung von XSLT-Skripten auf der Basis der GNU libxslt) und XML::XPath (API zur direkten Auswertung von XPath-Ausdrücken).

Für jedes aus der oben beschriebenen Kompilation resultierende Definitionsmuster wird ein XSLT-Skript generiert, das aus einem Dokument sämtliche `sent`-Elemente (also Sätze inklusive ihrer Parses) extrahiert, in deren PreDS mindestens ein `word`-Element (d.h. ein PreDS-Knoten) die *query* des jeweiligen Musters erfüllt, und jedes solche `word`-Element dann durch ein Attribut `definition='true'` markiert. Im Kontext dieser markierten Elemente werden dann die XPath-Ausdrücke für die Definitionsbestandteile ausgewertet und von den so ermittelten `word`-Elementen aus gegebenenfalls verschiedene vorab spezifizierte "transparente" Konstruktionen übersprungen (wie etwa *Begriff des…* (vgl. (5.2)), *Tatbestandsmerkmal…* oder *Vorliegen eines…*). Schließlich werden die Tokenpositionen abgelesen, die zu dem jeweiligen Definitionsbestandteil gehören und anhand derer dann die entsprechende Textspanne bestimmt werden kann.

Bei der beschriebenen Vorgehensweise muss für jedes Suchmuster jedes Dokument jeweils einmal geöffnet und seine XML-Struktur geparst werden. Auch wenn die Ausführung von XSL-Transformationen in der von uns eingesetzten libxslt sehr effizient implementiert ist, führt der mit diesen Vorgängen verbundene Overhead bei großen Korpora zu erheblichen Effizienzproblemen – mit der von uns genutzten Hardware nimmt die Durchsuchung des gesamten in Kap. 4 beschriebenen Korpus (33 579 Dokumente, ca. 40 Gigabyte XML-Daten) beispielsweise für den in Abb. 5.6 angegebenen Ausdruck 4,5 Stunden in Anspruch.

Da die Abarbeitung der Suchmuster sowie die Suche in den einzelnen Dokumenten jedoch vollständig unabhängig voneinander erfolgen, kann der Suchprozess feinkörnig verteilt werden. Für unsere Untersuchungen konnten wir die Auswirkungen der angesprochenen Effizienzprobleme daher durch parallele Verarbeitung minimieren. Steht endgültig fest, welche Information aus PreDS für die Definitionssuche relevant ist, lässt sich eine erhebliche Effizienzsteigerung ohne Parallelisierung voraussichtlich durch eine entsprechende selektive Abbildung des PreDS-Formats in ein relationales Datenbankschema erzielen. Werden PreDS-Knoten in einer Tabelle mit einer Stringrepräsentation des Pfades vom Wurzelknoten abgelegt, können die XPath-typischen flexiblen Adressierungsmöglichkeiten durch Matchingoperationen mit dem SQL LIKE-Operator ausgedrückt werden (etwa `descendant` als `PATH1 LIKE PATH2%`). In einem Experiment konnten wir mit einer entsprechenden manuellen SQL-

Übersetzung des Definitionsmusters in Abb. 5.6 und einer auf etwa 30% des CORTE-Korpus eingeschränkten Datenbank eine Reduktion der Laufzeit auf wenige Minuten feststellen. Für ein praktisch angewandtes System scheint es daher langfristig sinnvoll, eine relationale Datenbank und SQL-Anfragen anstelle der momentan von uns eingesetzten Lösung zu verwenden.

### 5.2.3 Genutzte Muster

Die linguistisch motivierte Grundstruktur des Spezifikationsformats erlaubt weitgehend eine direkte Umsetzung der Analysen aus 2.3 in dependenzbasierte Definitionsmuster. In Anhang C sind sämtliche 62 *key-frame*-Kombinationen angegeben, die zur Modellierung der Definitionsformulierungen aus unserem Entwicklungskorpus benötigt werden. Kopulabasierte Konstruktionen zur Definition prädikativ gebrauchter Adjektive und Partizipien (igs. je fünf Formulierungstypen in unseren Entwicklungsdaten) werden aufgrund von Besonderheiten des PreDS-Formats[5] durch *frames* ohne *key* spezifiziert. Insgesamt ergeben sich nach Kompilation des *key-frame-mapping*-Formats 76 ausführbare dependenzbasierte Suchmuster. Die gegenüber der sequenzbasierten Modellierung (93 Muster) erheblich geringere Musteranzahl ergibt sich vor allem dadurch, dass im PreDS-Format Passivumschreibungen und Stellungsvarianten trennbarer Verben nicht separat modelliert werden müssen.

Sudo u. a. (2003) und Stevenson und Greenwood (2006) charakterisieren verschiedene Typen dependenzbasierter IE-Modelle dadurch, welche Ausschnitte aus der Dependenzstruktur eines Satzes jeweils in Suchmustern verwendet werden. Stevenson und Greenwood (2006) unterscheiden zwischen (a) dem *Predicate-Argument*-Modell auf der Grundlage von Subjekt-Verb-Objekt-Tripeln, (b) dem *Chain*-Modell auf der Grundlage gerichteter Dependenzpfade von einem Verb, (c) dem *Linked-Chain*-Modell, das Muster aus Kombinationen mehrerer solcher Pfade erlaubt sowie (d) dem *Subtree*-Modell, in dem beliebige Ausschnitte aus einer Dependenzstruktur als Suchmuster dienen können. Nach dieser Einteilung können die von uns verwendeten Suchmuster als *Linked-Chain*-Muster mit einigen Erweiterungen (insbesondere der durch den `descendant`-Step in XPath realisierten Möglichkeit zur Verwendung von Wildcards in Pfaden) klassifiziert werden. Mit der Expressivität der IE-Modelle steigt in der angegebenen Reihenfolge (aufgrund der wachsenden Anzahl möglicher Konstellationen) auch die Komplexität der automatischen Akquisition

[5] In Adjektivprädikationen wird das Adjektiv statt der Kopula als Hauptprädikat repräsentiert. Wenn es sich bei diesem selbst um das Definiendum handelt, kann also kein allen entsprechenden Definitionen gemeinsames Prädikat im Suchmuster angegeben werden. Stattdessen wird im *frame* die in allen solchen Definitionen invariante Kategorie des Prädikats (*verb-adj*) spezifiziert.

entsprechender Suchmuster (vgl. Sudo u. a. (2003) und Stevenson und Greenwood (2006)). Wir gehen in Kap. 6 näher auf die Auswirkungen des von uns gewählten Modells im Hinblick auf den automatischen Mustererwerb ein.

## 5.3 Evaluation

Im Rest dieses Kapitels befassen wir uns mit der Evaluation der Definitionssuche im CORTE-System. Die Auswertung hat den praktischen Aspekt der Abschätzung der Ergebnisqualität, die von einer solchen Definitionssuche als Komponente eines juristischen Informationssystems zu erwarten ist (zu praktischen Anwendungskontexten vgl. Kap. 4 und Kap. 7). Aus theoretischer Sicht erlaubt sie zudem Rückschlüsse auf die Validität der Abstraktionen und Generalisierungen, die wir in Kap. 2 anhand von Korpusbeispielen getroffen haben.

### 5.3.1 Methodologische Anmerkungen

Bevor wir jedoch zur Präsentation der eigentlichen Evaluationsergebnisse kommen, möchten wir noch einige methodologische Fragen klären. Wir umreißen zunächst genauer die einzelnen Aufgabenstellungen, auf die wir bei der Evaluation Bezug nehmen und fassen noch einmal kurz bisherige Ansätze für die Evaluation von Systemen zur Definitionssuche zusammen (siehe hierzu auch Kap. 3). Dann gehen wir auf den Referenzdatenbestand und die Evaluationsmetriken ein, die wir verwenden werden.

#### (a) Aufgabenstellungen

Das im Rahmen der dependenzbasierten Suche genutzte PreDS-Format enthält genug Information, um eine Abbildung von Definitionsbestandteilen auf linguistische Analysekategorien zu ermöglichen. Es ermöglicht somit neben der Erkennung von Definitionen auch die Identifikation von Definitionsbestandteilen über Dependenzpfade vom Definitionsprädikat (siehe 5.2). Zwischen beiden Teilaufgaben (wir sprechen im Folgenden von *Definitionsextraktion* oder auch *-identifikation* und *Definitionssegmentierung*) besteht eine sequentielle Abhängigkeit, denn Definitionsbestandteile können nur in zuvor als Definition identifizierten Sätzen erkannt werden. Gleichzeitig ist die Qualität der Segmentierung korrekt identifizierter Definitionen aber unabhängig von der Qualität der Definitionsextraktion. Wir werden daher die beiden Teilaufgaben im Folgenden getrennt voneinander evaluieren und die Evaluation der Definitionssegmentierung auf korrekt identifizierte Definitionen einschränken. Definitionsextraktion und Definitionssegmentierung erfolgen im CORTE-System in zwei separaten

Schritten, so dass eine solche Trennung bei der Evaluation auch praktisch problemlos möglich war.

Im Falle der sequenzbasierten Suche steht kein erfolgversprechender Ansatzpunkt für die Definitionssegmentierung zur Verfügung, denn die benötigte Strukturinformation ist in reinen Lemmasequenzen nicht enthalten. Ergebnis des flachen Verarbeitungspfads im CORTE-System sind unsegmentierte Treffer. Die Evaluation für die sequenzbasierte Suche beschränkt sich daher auf die Definitionsextraktion.

#### (b) Evaluation von Systemen zur Definitionssuche

Wie in Kap. 3 diskutiert existieren für die Evaluation von Definitionsextraktionssystemen bisher keine Standardressourcen und keine übertragbaren Evaluationsschemata.

Für definitorische Question Answering-Systeme wurde im Rahmen der TREC- und CLEF-Wettbewerbe eine Evaluationsmethodologie entwickelt. Sie stützt sich auf Antwortschlüssel, die für eine Anzahl von Begriffen die wesentlichen Fakten spezifizieren, die in einer Definition enthalten sein müssen. Es wird dann ermittelt, wie viele der geforderten Fakten in den Systemantworten zu Definitionsfragen nach diesen Begriffen enthalten sind. Diese Vorgehensweise ermöglicht zwar eine *end-to-end*-Evaluation von QA-Systemen mit relativ geringem Aufwand. Sie ist jedoch stark auf die speziellen Anforderungen und Gegebenheiten des definitionsbasierten QA zugeschnitten (begriffszentrierter Zugriff auf Definitionen, knapp und vollständig definierbare Zielbegriffe, erhebliche Redundanzen in der Textgrundlage) und liefert bei verschiedenen Fragekatalogen keine direkt vergleichbaren Werte. Zudem kann sie nicht ohne weiteres für eine separate Evaluation einzelner Schritte bestimmter Verfahren, wie in unserem Fall der Definitionsextraktion und -segmentierung, verwendet werden.

Für die Evaluation von Verfahren zur direkten, begriffsunabhängigen Identifikation von Definitionen sind unseres Wissens bisher grundsätzlich individuelle Bezugskorpora mit manuell annotierten Definitionen verwendet worden (soweit überhaupt eine quantitative Evaluation vorgenommen wurde). Auf den Annotationsprozess wird nur selten näher eingegangen. Für die Bewertung der Übereinstimmung zwischen Extraktionsergebnissen und Annotation kommen die Standardmaße Präzision, Recall und f-Score zum Einsatz. Die Segmentierung extrahierter Definitionen ist in keinem uns bekannten Ansatz quantitativ ausgewertet worden.

**(c) Referenzdatenbestände**

Wir stützen uns bei der Evaluation von Definitionsextraktion und -segmentierung im CORTE-System ebenfalls in erster Linie auf individuell erzeugte Referenzressourcen, nämlich auf unser manuell annotiertes Pilotstudien- und Goldstandard-Korpus (vierzig bzw. sechzig Entscheidungstexte, vgl. Kap. 2), die wir im Folgenden auch als *Definitionskorpora* bezeichnen werden (siehe Anhang A für eine Übersicht über alle in dieser Arbeit erzeugten und genutzte Korpora).

Mit diesen Korpora verfügen wir über eine für die Rechtssprache einzigartige Ressource, die uns nicht nur die Bewertung der Gesamtperformanz unseres Suchverfahrens erlaubt, sondern auch eine detaillierte Fehleranalyse auf der Ebene einzelner *false positives* und *false negatives* (siehe 5.4). Für die Auswertung der Definitionssegmentierung wurde die in Kap. 2 beschriebene Annotation der Definitionskorpora um eine einfache Annotation von Definiens und Definiendum ergänzt. Diese wurde durch denselben juristischen Annotator (Jurist nach dem ersten Staatsexamen) vorgenommen, der auch an der Definitionsannotation beteiligt war.

Das *Goldstandard-Korpus* haben wir für Testzwecke zurückbehalten. Es stellt im Hinblick auf die Entwicklung des CORTE-Systems vollständig "ungesehenes" Material dar und ermöglicht somit eine Abschätzung der Ergebnisqualität bei der Verarbeitung neuer Datenbestände.

Das *Pilotstudien-Korpus* dagegen haben wir bereits bei der Suchmusterspezifikation als Entwicklungsdaten verwendet. Auf seiner Grundlage können deshalb nur mit Vorsicht Aussagen über die Leistungsfähigkeit des CORTE-Systems auf neuen, bei der Entwicklung nicht bekannten Texten getroffen werden. Es kann jedoch in jedem Fall eine obere Orientierungslinie für die Ergebnisqualität ermittelt werden. Anhand dieser kann abgeschätzt werden, in wie weit bei der Evaluation mit unbekannten Testdaten ermittelte Performanzprobleme auf die Veränderung der Datengrundlage zurückgehen, und in wie weit prinzipielle, datenunabhängige Probleme vorliegen. Im Hinblick auf Recall-Probleme erlaubt die Evaluation auf der Basis des Pilotstudien-Korpus dabei eine Einschätzung über die Auswirkungen von Problemen in der linguistischen Vorverarbeitung (Grammatik- und Programmfehler) und den Eingabedaten (z.B. Zeichenkodierungsfehler). Im Hinblick auf Präzisionsprobleme können wir noch weitreichendere Rückschlüsse ziehen: Wir haben die hier betrachteten Suchmuster vollständig anhand von Positivbeispielen aus unseren Entwicklungsdaten modelliert. Über den Abstraktions- bzw. Einschränkungsgrad der einzelnen Muster haben wir im wesentlichen aufgrund genereller Erwägungen entschieden. Der nicht-definitorische Anteil der Entwicklungsdaten bleibt daher weiterhin "neu": Selbst im Falle einer optimalen Vorverarbeitung

könnte nur empirisch ermittelt werden, wie viele *false positives* unsere Suchmuster aus den Entwicklungsdaten extrahieren.

Zusätzlich können wir für einige Analysen auf das *CORTE-Großkorpus* zurückgreifen (33 579 Entscheidungstexte, genauere Angaben zum Aufbau finden sich im vorigen Kapitel, 4.3.1) und das *juris-Definitionsregister* mit einbeziehen (vgl. 2.4.2). Im *CORTE-Großkorpus* sind zwar keine Definitionen annotiert, denn hierfür wäre die vollständige Durchsicht aller enthaltenen Entscheidungstexte notwendig gewesen. Die Analyse der Extraktionsergebnisse erlaubt uns aber dennoch zumindest die Bestimmung der Präzision der Definitionsextraktion. Diese Maßzahl können wir somit zusätzlich noch auf einer erheblich größeren Datengrundlage bestimmen, als sie mit dem Goldstandard- und dem Pilotstudien-Korpus verfügbar ist.

Da sowohl für das Goldstandard-Korpus als auch für einen großen Teil der Entscheidungen aus dem CORTE-Großkorpus Einträge im *juris-Definitionsregister* vorhanden sind, werden wir die Ergebnisse unserer Definitionsextraktion und -segmentierung schließlich zusätzlich durch einen Vergleich mit dieser von Domänenexperten manuell erstellten Ressource validieren.

#### (d) Evaluationsmetriken

Als Evaluationsmetriken verwenden wir die Maßzahlen Präzision (p), Recall (r) und f-Score (f). Ihre Definitionen geben wir im Folgenden noch einmal wieder (vgl. auch 3.1.1):

$$p = \frac{|\text{Korrekte Treffer}|}{|\text{Alle Treffer}|}$$

$$r = \frac{|\text{Korrekte Treffer}|}{|\text{Geforderte Treffer}|}$$

$$f_\beta = (1+\beta^2)\frac{pr}{\beta^2 p + r}$$

Der in der Definition des f-Score enthaltene Parameter $\beta$ läßt eine Gewichtung zwischen Präzision und Recall zu. Je nach Anwendungskontext können hier verschiedene Werte sinnvoll sein. So dürfte zum Beispiel bei einer Websuche aufgrund der hohen Redundanz der Information des WWW ein relativ niedriger Recall eher zu tolerieren sein als eine geringe Präzision (und somit eine große Anzahl irrelevanter Treffer). Ein f-Score mit $\beta \neq 1$ kann auch berechnet werden, um bekannten Möglichkeiten einer späteren Ergebnisoptimierung Rechnung zu tragen. Wir wenden uns den Themen Optimierungsmöglichkeiten und Anwendungen in Kap. 6 bzw. Kap. 7 zu. Hier machen wir zur Gewichtung von Präzision und Recall vorerst keine Annahmen. Im Folgenden geben wir

f-Scores an, die mit $\beta = 1$, also als harmonischer Mittelwert von Präzision und Recall, berechnet sind.

Für die Ermittlung von Präzision und Recall nach obigen Definitionen ist entscheidend, wie die Größen in Zähler und Nenner definiert werden. Es muß deshalb geklärt werden, welche Granularität (und somit welche *Anzahl geforderter Treffer*) betrachtet wird, was als Treffer gewertet wird und wann ein Treffer als korrekt gilt. Die Antworten auf diese Fragen hängen von der untersuchten Aufgabenstellung ab. Wir erläutern im Folgenden, wie wir in diesen Punkten bei der Evaluation der Definitionsextraktion und -segmentierung im CORTE-System vorgegangen sind.

**Definitionsextraktion.** Als Bewertungseinheit im Falle des Definitionsextraktionsschritts haben wir Sätze gewählt. Mehrfach extrahierte Sätze wurden nur einmal gewertet, und Sätze mussten nur einmal extrahiert werden, auch wenn sie mehrere Definitionen enthielten.

Außerdem war für die Ermittlung von Präzision und Recall relativ zu unseren annotierten Definitionskorpora zu entscheiden, wie mit der verhältnismäßig reichhaltigen annotierten Information im Einzelnen umgegangen werden sollte. Bei der Erstellung der Suchmuster für das CORTE-System wurden nur prädikatbasierte Formulierungen berücksichtigt (vgl. 2.3.2). Im Hinblick auf die Unterscheidung zwischen *prädikatbasierter* und *appositiv-parenthetischer* Definitionsrealisierung stand daher fest, dass sinnvollerweise nur prädikatbasierte Definitionen als geforderte Treffer für die Recall-Berechnung zu Grunde gelegt werden konnten. Ebenso waren beide annotierten Informationstypen (*Kern- vs. elaborierende Information*) als Treffer zu fordern, da die Instanzen beider Typen in Suchmuster umgesetzt wurden. Darüber hinaus wurden in drei weiteren Punkten Festlegungen benötigt:

- zum Umgang mit den verschiedenen Konfidenzstufen, die bei der Annotation vergeben wurden (*klar*, *unklar* und *zweifelhaft*)
- für die Recall-Berechnung zur Bewertung von Fällen, in denen nur ein Teil der Sätze aus einem mehrsätzigen Definitionskomplex extrahiert wurde (bei der Annotation konnten mehrere Sätze zu einem Definitionskomplex zusammengefasst werden)
- für die Präzisionsberechnung zur Bewertung von (nicht geforderten) "Zufallstreffern" in appositiv-parenthetischen Definitionen

Wir haben für die Berechnung von Präzision und Recall die folgenden drei Rahmenbedingungen definiert:

(a) *Eng*: Es werden Treffer in den Sätzen gefordert und als korrekt gewertet, die als *Definitionskern* oder *elaborierende Information* annotiert sind, bei denen es sich um prädikatbasierte Konstruktionen handelt und die zu einer Definition gehören, der der Annotator den Konfidenzwert *klar* zugewiesen hat. In diesem Szenario wurden im Pilotstudien-Korpus 95 und im Goldstandard-Korpus 275 Treffer gefordert.

(b) *Erweitert 1*: Es werden auch Treffer als korrekt gewertet und gefordert, die in als *unklare* Fälle klassifizierten Definitionen liegen. Potentiell erhöht sich hierdurch die Präzision (mehr Treffer werden als korrekt gewertet), während der Recall sinkt (da mehr Sätze gefunden werden müssen). Die Anzahl der geforderten Treffer in diesem Szenario betrug 135 für das Pilotstudien- und 426 für das Goldstandard-Korpus.

(c) *Erweitert 2*: Ergänzend zu den Erweiterungen in (b) werden Recall-Werte auf Definitionsebene berechnet. Eine (u.U. mehrsätzige) Definition wird dabei als gefunden gewertet, wenn mindestens ein zu ihr gehörender Satz extrahiert wurde. Dies führt potentiell zu einer Verbesserung der Recall-Werte. Für die Ermittlung der Präzisionswerte werden zudem auch Treffer als korrekt gewertet, die in appositiven Definitionen liegen. Hierdurch erhöht sich potentiell der berechnete Präzisionswert.[6]

Das Szenario *Erweitert 2* beinhaltet aus praktischer Sicht motivierte Abschwächungen: Einzelsätze aus komplexen Definitionen können einen Einstiegspunkt bilden, von dem aus die gesamte Definition erschlossen werden kann (im einfachsten Fall durch Extraktion des Kontexts, bessere Ergebnisse würde wohl ein gezieltes Verfolgen von Diskursrelationen liefern). Appositive Definitionen dürften in den meisten Anwendungskontexten als Treffer brauchbar sein, unabhängig davon, ob eigentlich nur prädikatbasierte Definitionen erwartet wurden.

Die Präzisionswerte für das CORTE-Großkorpus wurden ermittelt, indem eine zufällige, jedoch über die einzelnen Suchmuster gleichmäßig verteilte Auswahl von je ca. 6000 Treffern (von den insgesamt 1 044 663 bzw. 108 209 Treffern der sequenz- bzw. dependenzbasierten Suchmuster) in einem vereinfachten Verfahren ausgewertet wurden. Sie waren ohne Angabe von Konfidenzwerten und Informationstypen als *definitorisch* oder *nicht-definitorisch* zu klassifizieren. Da zudem nur Einzelsätze bewertet wurden, entfällt hier die Unterscheidung zwischen den drei genannten Evaluationsszenarien.

[6] Diese Erweiterung ist vorwiegend für den Fall der wenig trennscharfen sequenzbasierten Suchmuster relevant. Im Fall der dependenzbasierten Extraktion aus dem Goldstandard-Korpus waren dagegen z.B. nur zwei Definitionen betroffen.

Die Annotation erfolgte durch den juristischen Annotator. Zur Überprüfung wurden je ca. 2000 der ausgewählten sequenz- bzw. dependenzbasierten Treffer zusätzlich vom Autor dieser Arbeit bewertet. Die Korrelation betrug $\kappa$=0,83 für die dependenzbasierten und $\kappa$=0,85 für die sequenzbasierten Treffer.[7]

**Definitionssegmentierung.** Für die Auswertung der Ergebnisse der Definitionssegmentierung musste zunächst einmal festgelegt werden, welche Granularität bei der Evaluation zu Grunde liegen sollte. Ein Vergleich auf der Ebene vollständiger Definitionsbestandteile (also jeweils des kompletten extrahierten und annotierten Definiens und Definiendum) erscheint zunächst als naheliegende Wahl. Diese Methode kann jedoch keine Segmentierungsergebnisse berücksichtigen, die sich mit den realen Definitionsbestandteilen zwar überlappen, diese aber nur unvollständig abdecken oder noch weiteres Material umfassen. Den ersten Fall illustriert (5.3), den zweiten (5.4) (unterstrichen ist das annotierte Definiendum, fett gedruckt das von der Segmentierung erkannte).

(5.3) Um "wissenschaftlich" **anerkannt** zu sein, müssen Beurteilungen von solchen Personen vorliegen, die an Hochschulen und anderen Forschungseinrichtungen als Wissenschaftler in der jeweiligen medizinischen Fachrichtung tätig sind.

(Verwaltungsgerichtshof Baden-Württemberg 4. Senat, 3. Mai 2002, AZ 4 S 512/02, juris)

(5.4) Insoweit setzt **die Feststellung einer Verunstaltung** kein so krasses geschmackliches Unwerturteil wie das einer das ästhetische Empfinden verletzenden Häßlichkeit voraus.

(Oberverwaltungsgericht Berlin 2. Senat, 31. Juli 1992, AZ 2 B 14.90, juris)

Je nach Grad der Überlappung können solche Ergebnisse jedoch durchaus brauchbar sein. Wir haben uns daher entschieden, Präzision und Recall auf Wortebene zu berechnen. Geforderte Treffer waren die Mengen der Tokens in den als Definiens bzw. Definiendum annotierten Textspannen. Als erzielte Treffer galten die Tokenmengen aus den Textspannen, die von der Segmentierung als Definiens bzw. Definiendum erkannt wurden. Als korrekte Treffer wurde die Schnittmenge aus erzielten Treffern und geforderten Treffern gezählt.

[7] Sie liegt somit deutlich über dem für den Goldstandard ermittelten Wert ($\kappa$=0,58). Ein Hauptgrund hierfür dürfte darin liegen, dass nur eine einzelne binäre Unterscheidung zu treffen war. Zudem waren Definitionen in dem hier betrachteten Annotationsszenario nicht im Fließtext zu identifizieren, sondern in einer – als Ergebnis des Extraktionsverfahrens – sprachlich sehr viel einheitlicheren Vorauswahl mit einem gegenüber dem Fließtext insgesamt höheren Anteil eindeutiger (nämlich eindeutig positiver) Fälle.

### 5.3.2 Definitionsextraktion

Wir betrachten nun die Qualität der Ergebnisse der Definitionsextraktion aus Pilotstudien-, Goldstandard- sowie dem gesamten CORTE-Korpus. Die Segmentierungsergebnisse evaluieren wir in 5.3.3. Eine Auswertung gegen das *juris*-Definitionsregister nehmen wir in 5.3.4 vor.

#### (a) Ergebnisse

In Tabelle 5.2 sind Präzision, Recall und f-Score für die sequenz- sowie dependenzbasierte Extraktion aus unseren Korpora aufgeführt. Insgesamt wurden durch die sequenzbasierte Suche im Pilotstudien-Korpus 1837 Sätze und im Goldstandard-Korpus 4196 Sätze extrahiert. Die dependenzbasierten Suchmuster lieferten Treffer in 296 bzw. 616 verschiedenen Sätzen.

Die verwendeten Evaluationsbedingungen *Eng*, *Erweitert 1* und *Erweitert 2* sind im vorigen Abschnitt unter (d) näher erläutert. Für das CORTE-Großkorpus konnten wie bereits angesprochen keine Recall-Werte bestimmt werden, da in diesem Korpus keine Definitionen annotiert sind. Bei den in Klammern angegebenen Werten handelt es sich um Schätzungen. Sie basieren auf der Annahme, dass die über Pilotstudien und Goldstandard-Korpus gemittelte "Definitionsquote" von 3,6% (prädikatbasierte definitorische Sätze in Leit- / Orientierungssätzen und Entscheidungsgründen) auf das Gesamtkorpus verallgemeinert werden kann. Es wären dort somit ca. 59 200 solche Sätze zu erwarten. Diese Werte sind natürlich aufgrund des Übergangs von einer kleinen Grundgesamtheit auf eine sehr große Zahl mit Vorsicht zu betrachten und können nur eine recht grobe Orientierung geben.

#### (b) Analyse und Bewertung

Die ermittelten Werte bestätigen klar die Vermutung, dass durch den Zugriff auf Dependenzinformation eine erheblich präzisere Definitionserkennung möglich ist als auf der Basis von Wort- bzw. Lemmasequenzen. Mit diesem Vorteil geht zwar ein geringerer Recall einher (dessen genaue Ursachen wir in 5.4 näher untersuchen werden). Dieser wird jedoch in den f-Scores durchgängig durch die erhöhte Präzision aufgewogen.

Die auf fremden Daten (Goldstandard- und CORTE-Korpus) erzielten Werte lassen es realistisch erscheinen, dass die dependenzbasierte Suche für Aufgaben, die eine manuelle Nachbearbeitung einschließen, direkt einsetzbar wäre, also etwa als automatischer Teilschritt bei der Verschlagwortung oder zur automatischen Erzeugung von Leitsatz-Kandidaten bei der dokumentarischen Erfassung von Entscheidungstexten. Für eine Nutzung zum vollautomatischen

| | sequenzbasiert | | | dependenzbasiert | | |
|---|---|---|---|---|---|---|
| | *p* | *r* | *f* | *p* | *r* | *f* |
| **Pilotstudien-Korpus (Entwicklungsdaten)** | | | | | | |
| *(a) Eng* | 0,05 | 0,94 | 0,09 | 0,20 | 0,62 | 0,30 |
| *(b) Erweitert 1* | 0,07 | 0,93 | 0,13 | 0,27 | 0,59 | 0,37 |
| *(c) Erweitert 2* | 0,08 | 0,98 | 0,14 | 0,28 | 0,69 | 0,40 |
| **Goldstandard-Korpus (Testdaten)** | | | | | | |
| *(a) Eng* | 0,05 | 0,70 | 0,09 | 0,13 | 0,28 | 0,17 |
| *(b) Erweitert 1* | 0,07 | 0,69 | 0,13 | 0,19 | 0,27 | 0,22 |
| *(c) Erweitert 2* | 0,07 | 0,89 | 0,14 | 0,19 | 0,45 | 0,27 |
| **CORTE-Großkorpus (Testdaten)** | | | | | | |
| *Trefferanalyse* | 0,04 | *(0,75)* | *(0,08)* | 0,20 | *(0,39)* | *(0,26)* |

Tabelle 5.2: Ergebnisse der Definitionsextraktion

Zugriff auf Definitionen, zum Beispiel in einer juristischen Suchmaschine, sind allerdings weitere Optimierungen wünschenswert.

Die sequenzbasierte Extraktion erreicht zwar auf den Entwicklungsdaten nahezu vollständigen Recall, und auch auf dem Goldstandard-Korpus werden noch sehr hohe Recall-Werte erzielt. Jedoch liegen die Präzisionswerte mit durchgängig unter 10% in einem Bereich, der, wenn überhaupt, nur in sehr wenigen Kontexten (z.B. für eine grobe Filterung als Vorstufe bei der manuellen Textzusammenfassung) anwendungstauglich sein dürfte. Auch als Fallback-Verfahren für die dependenzbasierte Suche erscheint die sequenzbasierte Suche nur unter der Bedingung geeignet, dass die Ergebnisqualität durch nachträgliche Optimierung verbessert werden kann.

Für beide Suchmustergruppen führt die Abschwächung der Evaluationsbedingungen im Szenario *Erweitert 1* zu höheren Präzisionswerten. In *Erweitert 2* ist durch die positive Bewertung von Treffern in appositiven Definitionen nur im Pilotstudien-Korpus eine leichte Präzisionsverbesserung zu verzeichnen. Die Recall-Werte erhöhen sich dagegen durch die definitionsbasierte Recall-Berechnung durchgängig deutlich.

Der Übergang von Entwicklungs- zu Testdaten zieht zwar insgesamt einen Performanzverlust nach sich (der im Fall der dependenzbasierten Extraktion

deutlicher ausfällt als bei der sequenzbasierten Suche). Tabelle 5.2 ist allerdings zu entnehmen, dass dieser in höherem Maße auf eine Verringerung des Recall zurückzuführen ist als auf die Präzisionswerte. Eine detailliertere Betrachtung der Präzision auf Suchmusterebene zeigt interessanterweise, dass sich für einen gewissen Teil der Muster auf den Testdaten sogar Präzisionsverbesserungen ergeben. Diese Ergebnisse deuten darauf hin, dass die im Rahmen unserer Pilotstudie ermittelten definitorischen Formulierungsmuster zwar zu einem Teil relativ gut auf den allgemeinen Fall übertragen werden können. Der gegenüber der Extraktion aus den Entwicklungsdaten erheblich verringerte Recall lässt jedoch erkennen, dass die ermittelten Muster bei weitem nicht das gesamte gängige Inventar solcher Ausdrücke widerspiegeln.

Eine Analyse der Performanz der einzelnen Suchmuster hat gezeigt, dass für beide Suchmechanismen im Hinblick auf die Präzision offensichtliches Optimierungspotential gegeben ist. Ein gewisser Teil der Treffer im Goldstandard-Korpus (ca. 10% für die dependenzbasierte und ca. 5% für die sequenzbasierte Extraktion) werden von Suchmustern mit einer Präzision von 0,5 oder mehr erzielt. Gleichzeitig entfällt ein nicht unerheblicher Teil der Treffer (etwa ein Fünftel bei der dependenzbasierten und gut zwei Drittel bei der sequenzbasierten Extraktion) auf Suchmuster mit einer Präzision von höchstens 0,2. Allein durch eine geeignete Suchmusterauswahl dürfte also eine Erhöhung der Präzision ohne allzu große Recall-Verluste erzielbar sein.

Für eine Erhöhung des Recall hingegen sind zum einen weitere Verbesserungen in der Vorverarbeitung notwendig. Schon die in Kap. 4 beschriebenen Anpassungen des PreDS-Parsers haben hier einen erkennbaren Fortschritt erbracht. Der Recall bei der dependenzbasierten Extraktion aus einer mit PreDS.GF (also ohne unsere Optimierungen) geparsten Version des Pilotstudien-Korpus liegt mit 0,61 (für Szenario *Erweitert 2*) deutlich unter dem in Tabelle 5.2 genannten Wert. Zum anderen werden weitere Suchmuster benötigt.

In Kap. 6 werden wir uns mit Methoden zur Optimierung von Präzision und Recall der dependenzbasierten Extraktion u.a. auf der Basis der genannten Ansätze befassen. Wir gehen davon aus, dass ein großer Teil der dort erprobten Verfahren in ähnlicher Weise auch für die sequenzbasierte Extraktion anwendbar sind, können diese Frage aber im Rahmen unserer Arbeit nicht mehr vertiefen.

### 5.3.3 Definitionssegmentierung

#### (a) Ergebnisse

Tabelle 5.3 sind Präzision, Recall und f-Scores für die Segmentierung der korrekten Definitionen zu entnehmen, die durch die dependenzbasierten Suchmu-

ster aus Pilotstudien- und Goldstandard-Korpus extrahiert wurden. Die Werte sind wie in 5.3.1 unter (d) beschrieben auf der Basis von Tokenmengen ermittelt. Neben den tatsächlich erzielten Werten sind zum Vergleich jeweils die Erwartungswerte für eine zufällige Auswahl einer entsprechenden Anzahl von Tokens angegeben.

| | *p* | *r* | *f* |
|---|---|---|---|
| **Pilotstudien-Korpus (Entwicklungsdaten)** | | | |
| *Alle Bestandteile* | 0,87 (0,61) | 0,64 (0,44) | 0,74 (0,51) |
| *Definiendum* | 0,46 (0,08) | 0,63 (0,11) | 0,53 (0,10) |
| *Definiens* | 0,92 (0,54) | 0,61 (0,36) | 0,73 (0,43) |
| **Goldstandard-Korpus (Testdaten)** | | | |
| *Alle Bestandteile* | 0,71 (0,57) | 0,54 (0,43) | 0,61 (0,49) |
| *Definiendum* | 0,23 (0,07) | 0,41 (0,12) | 0,30 (0,09) |
| *Definiens* | 0,61 (0,51) | 0,38 (0,33) | 0,47 (0,40) |

Tabelle 5.3: Performanz der Erkennung von Definitionsbestandteilen (Erwartungswerte in Klammern)

#### (b) Analyse und Bewertung

Auf dem Pilotstudien-Korpus werden für die Definitionssegmentierung relativ hohe Scores erzielt. Dies zeigt, dass mit der Verfolgung von Dependenzen des Definitionsprädikats ein zuverlässiger Mechanismus für die Identifikation von Definitionsbestandteilen zur Verfügung steht. Probleme der Vorverarbeitung stellen den Hauptfaktor dar, der die Segmentierungsqualität auf den Entwicklungsdaten noch einschränkt.

Die relativ niedrige Präzision bei der Erkennung von Definienda ist vor allem durch die insgesamt relativ geringe Anzahl von Tokens in diesem Segment bedingt (die sich auch in dem niedrigen Erwartungswert für eine "zufällige Segmentierung" niederschlägt). Hier fällt ins Gewicht, dass unsere Segmentierungsregeln nur einen Teil der vorkommenden *transparenten Konstruktionen* (wie *die Annahme eines…*) berücksichtigen und solches Material ansonsten fälschlicherweise dem Definiendum zugeschlagen wird. Dasselbe gilt für eingebettete Konstruktionen im Definiendum (die, wie im Fall von Relativsätzen, sehr umfangreich sein können).

Auch auf den Testdaten liefert die Definitionssegmentierung insgesamt akzeptable Ergebnisse. Selbst der Präzisionswert von 0,23 für die Definiendum-

Erkennung dürfte für Anwendungen mit manueller Nachbearbeitung (etwa für eine Unterstützung bei der Verschlagwortung) noch ausreichen. In anderen Fällen (etwa wenn aus erkannten und segmentierten Definitionen automatisch eine Domänenontologie aufgebaut oder Schlussfolgerungen gezogen werden sollen), wird allerdings eine nachträgliche Optimierung erforderlich sein. Wir werden die Frage, wie ein solcher Schritt zu realisieren wäre, hier nicht mehr weiter verfolgen können. Ein Ansatzpunkt wäre zum Beispiel die Verwendung statistischer Termgewichtungsmaße oder linguistisch motivierter Termformregeln. Diese würden es erlauben, innerhalb von Textspannen, die als Definiendum erkannt wurden, einen relevanten Kernbereich einzugrenzen. Da in über der Hälfte aller Definitionen die Definiendum-Token mit einem f-Score von mehr als 0,8 identifiziert werden können, erscheint es außerdem lohnend, nach Möglichkeiten zur Vorsortierung der Treffer der Definitionsextraktion in zuverlässig und weniger zuverlässig segmentierbare Definitionen zu suchen. Ein erstes Kriterium könnte hier (wie bei der Definitionsextraktion) das Suchmuster sein, mit dem eine Treffer erzielt wurde, denn für mehr als ein Drittel der Suchmuster liegt der f-Score bei der Definiendum-Erkennung ebenfalls über 0,8.

### 5.3.4 Vergleich mit dem *juris*-Definitionsregister

Abschließend wollen wir nun noch das Definitionsregister der Firma *juris* (vgl. 2.4.2) als Referenz zur Bewertung der Ergebnisqualität unseres Definitionsextraktionsverfahrens heranziehen. Das Register verzeichnet als Indextermini zu einem Teil der bei *juris* verfügbaren Entscheidungen ausgewählte Begriffe, die im Text definiert werden. Es kann in einem erweiterten Suchmodus über das Webportal der Firma gezielt durchsucht werden.

#### (a) Eigenschaften des *juris*-Definitionsregisters

Das Definitionsregister deckt nicht nur das Goldstandard-Korpus ab, sondern auch etwa ein Viertel der in CORTE-Großkorpus enthaltenen Entscheidungstexte (8509 der insgesamt 33 579 Dokumente). Es eignet sich somit als Basis für eine Abschätzung des Recall auf dieser Textgrundlage. Somit dient es als Ergänzung zu der stichprobenbasierten Präzisionsschätzung, deren Ergebnisse wir in 5.3.2 dargestellt haben.

Die Autoren des Definitionsregisters sind juristische Dokumentare, d.h. geschulte Domänenexperten. Es kann also insgesamt von einer hohen inhaltlichen Qualität der Registereinträge ausgegangen werden. Wie in 2.4.2 erläutert wird das Register jedoch pragmatisch und ohne feste Richtlinien geführt.

Zum einen deckt es oft nicht alle in einem Urteil auftretenden Definitionen ab – offenbar wird von den Dokumentaren bei der Verschlagwortung eine im-

plizite Relevanzgewichtung vorgenommen (vgl. zum Verhältnis zwischen Definitionsregister und unserer Annotation im Goldstandard-Korpus Tabelle 2.12 in 2.4.2). Zum anderen sind die als Indextermini fungierenden Ausdrücke formal uneinheitlich, nicht konsequent normalisiert und nicht auf eine automatische Verarbeitung hin formuliert. Insgesamt führt eine Vielzahl von zusammenfassenden Umschreibungen, Umstellungen und morphologischen Umformungen dazu, dass die Indextermini mancher Registereinträge textuell in dem indizierten Dokument gar nicht im Wortlaut auftreten. So finden sich neben Fällen, in denen komplexe Ausdrücke vollständig aus dem Urteilstext ins Register übernommen wurden (z.B. *sich dem Antritt der erkannten Freiheitsstrafe entziehen*) auch solche, in denen jeder einzelne Termbestandteil in einer Grundform als eigenes Schlagwort verzeichnet wurde (z.B. *gleichartig* und *Grundsatz* für *Grundsatz der Gleichartigkeit*).

#### (b) Vorgehensweise

Für die Evaluierung bedeutet dies zunächst einmal, dass das Register nicht als Präzisionsmaßstab verwendet werden kann. Es ist davon auszugehen, dass bei Weitem nicht zu allen brauchbaren extrahierten Definitionen tatsächlich ein Registereintrag vermerkt worden ist. Desweiteren kann beim Abgleich unserer Extraktionsergebnisse mit dem Definitionsregister kein vollständige textuelle Übereinstimmung mit den einzelnen Registereinträgen erwartet werden.

Wir untersuchen im Folgenden, für wie viele der Registereinträge eine zufriedenstellende Übereinstimmung mit mindestens einem Extraktionsergebnis aus demselben Urteil besteht. Wir betrachten dabei sowohl die extrahierten Definitionen insgesamt als auch die identifizierten Definienda. Um Normalisierungen und Modifikationen möglichst weitgehend zu kompensieren, ermitteln wir die Übereinstimmung auf der Grundlage eines Lemmamengen-Vergleichs zwischen Registereinträgen und Extraktionsergebnissen. Das heißt zum Beispiel, dass wir den oben genannten Indexterm *sich dem Antritt der erkannten Freiheitsstrafe entziehen* und ein extrahiertes Definiendum *entzieht sich dem Antritt der erkannten Freiheitsstrafe* als übereinstimmend betrachten. Normalisierungen wie *Gleichartigkeit* $\rightarrow$ *gleichartig* können allerdings nicht nachvollzogen werden. Auch Definitionen mit anaphorischem (also in der Definition selber als Pronomen realisiertem) Definiendum wirken sich in der Bewertung natürlich zu Ungunsten des CORTE-Systems aus.

#### (c) Ergebnisse

In Tabelle 5.4 ist für Goldstandard- und CORTE-Großkorpus angegeben, für wie viele Registereinträge vom CORTE-System jeweils passende Definitionen

gefunden werden konnten. Die Ergebnisse der sequenzbasierten Suche haben wir aufgrund der geringen Gesamtpräzision und der fehlenden Definitionssegmentierung außer Betracht gelassen. Es sind drei Szenarios mit verschiedenen Varianten der beschriebenen Vergleichsmethode unterschieden. Ein Registereintrag gilt als gefunden, wenn

(a) mindestens eine extrahierte Definition aus dem entsprechenden Urteil mindestens 75% der Lemmata des Indexterms enthält;

(b) mindestens ein extrahiertes Definiendum aus dem entsprechenden Urteil mindestens 50% der Lemmata des Indexterms enthält;

(c) außerdem mindestens 50% der Lemmata aus dem extrahierten Definiendum auch im Indexterm enthalten sind.

Die Schwellenwerte für die Überschneidung sollen Normalisierungen und Modifikationen in den Indextermen auffangen, die durch den lemmabasierten Vergleich nicht kompensiert werden können. Zusätzlich gleichen sie Abweichungen aus, die entstehen, weil die Lemmatisierung der Extraktionsergebnisse und der Indexterme aus dem Definitionsregister mit unterschiedlichen Werkzeugen durchgeführt wurden.[8] Die niedrigeren Schwellenwerte in den Szenarios (b) und (c) tragen der Tatsache Rechnung, dass Definienda oft nur wenige Token umfassen (vgl. die Erwartungswerte in Tabelle 5.3).

| | *Goldstandard* | *CORTE-Großkorpus* |
|---|---|---|
| *Registereinträge insgesamt* | 91 | 10409 |
| Davon gefunden: | | |
| *Szenario (a)* | 65 | 5537 |
| *Szenario (b)* | 40 | 3339 |
| *Szenario (c)* | 24 | 1556 |

Tabelle 5.4: Abdeckung des *juris*-Definitionsregisters durch Definitionsextraktionsergebnisse

In Szenario (a), also unter Einbeziehung aller Lemmata in den Treffern, wird im Goldstandard-Korpus für über zwei Drittel und im Großkorpus für mehr als die Hälfte der Registereinträge eine Definition im entsprechenden Urteil gefunden. Sowohl im diesem Szenario als auch bei der Einschränkung auf Definienda

[8] Für die Lemmatisierung des Definitionsregisters haben wir TreeTagger verwendet (vgl. 4.2.2). Die Lemmatisierung der Extraktionsergebnisse wurde der PreDS-Analyse entnommen und basiert auf Gertwol (siehe 4.2.3).

in Szenario (b) (wo zusätzlich zur Definitionsextraktion auch die Qualität Definitionssegmentierung relevant wird) dürfte der Recall hinreichend sein, um beispielsweise automatisch Vorschläge für Einträge in das Definitionsregister zu generieren. Der deutlich geringere Recall bei Hinzunahme der Präzisionsanforderung in Szenario (c) korreliert mit den recht niedrigen Präzisionswerten bei der Definiendum-Erkennung, die wir im vorigen Abschnitt festgestellt haben. Wir gehen davon aus, dass auch hier Verbesserungspotential in zusätzlichen Regeln zur Entfernung von transparentem Material und eingebetteten Konstruktionen aus den erkannten Definienda liegt.

## 5.4 Fehleranalyse

Um die Ursachen für Präzisions- und Recall-Fehler des CORTE-Systems genauer diagnostizieren zu können, haben wir eine Untersuchung von *false positives* (also inkorrekten Treffern) und *false negatives* (also nicht gefundenen Definitionen) des Definitionsextraktionsschritts vorgenommen. Wir erläutern in diesem Abschnitt zunächst unsere Vorgehensweise bei dieser Fehleranalyse und diskutieren dann die Ergebnisse.

### 5.4.1 Vorgehensweise

Konkrete Rückschlüsse auf die Gründe von Präzisions- und Recall-Problemen erfordern die Betrachtung einzelner Fehlerinstanzen. Im Rahmen unserer Fehleranalyse haben wir daher individuelle Präzisionsfehler (*false positives*) und Recall-Fehler (*false negatives*) der sequenz- und dependenzbasierten Extraktion aus unseren Definitionskorpora genau auf ihre Ursachen hin untersucht. Aufgrund des hiermit verbundenen Arbeitsaufwands konnten wir nur für einen Teil der Kombinationen von Fehlertyp und Datengrundlage alle Instanzen untersuchen. Für das Goldstandard-Korpus sowie für die Präzisionsfehler der sequenzbasierten Extraktion aus dem Pilotstudien-Korpus mussten wir uns auf Stichproben beschränken. Diese umfassen jeweils 70 Recall-Fehler bzw. – da eine gleichmäßige Verteilung über die Suchmuster garantiert werden sollte – etwa 200 Präzisionsfehler.[9] Die genaue Anzahl der untersuchten *false positives* und *false negatives* für die betrachteten Konstellationen ist in Tabelle 5.5 aufgeschlüsselt. Die Fehlerinstanzen wurden gemäß dem Szenario *Erweitert 1*

[9] Im Fall der Präzisionsfehler wurden die Stichproben erzeugt, indem so oft für alle Suchmuster ohne Zurücklegen ein *false positive* gezogen wurde, bis der Stichprobenumfang erstmals über 200 lag. Pro Suchmuster wurden so (je nach Konstellation) zwischen vier und sechs Fehlerinstanzen berücksichtigt.

ermittelt (siehe 5.3.2). Die Zahlen für *false positives* beziehen sich auf Satz-Suchmuster Kombinationen, da derselbe Satz als inkorrekter Treffer verschiedener Suchmuster auftreten kann.[10]

| | *Recall-Fehler* | *Präzisionsfehler* |
|---|---|---|
| **Pilotstudien-Korpus** | | |
| *sequenzbasiert* | 10 | 236 |
| *dependenzbasiert* | 56 | 228 |
| **Goldstandard-Korpus** | | |
| *sequenzbasiert* | 70 | 261 |
| *dependenzbasiert* | 70 | 207 |

Tabelle 5.5: Anzahl der untersuchten Fehlerinstanzen

### 5.4.2 Fehlerklassen

Aufbauend auf den Proben der *false positives* und *false negatives* haben wir eine genauere Unterteilung der beiden Hauptkategorien *Recall-Fehler* und *Präzisionsfehler* vorgenommen. In einem zweiten Durchgang haben wir dann die betrachteten Fehlerinstanzen den Unterkategorien zugeordnet. Wir erläutern nun zunächst qualitativ die erarbeiteten Fehlerklassen und diskutieren dann die quantitativen Ergebnisse unserer Fehleranalyse.

Die aufgrund der untersuchten Fehlerinstanzen ermittelten Unterklassen von *Recall-* und *Präzisionsfehlern* sind in Tabelle 5.6 zusammengefasst. Eine ausführlichere Beschreibung findet sich in Anhang D.

#### (a) A1–A5 bzw. B1: Vorverarbeitungsfehler

In beiden Hauptkategorien können zunächst einmal Vorverarbeitungsprobleme (A1–A5 bzw. B1) von anderen Fehlerursachen unterschieden werden. Schon die erste Durchsicht der Daten hat gezeigt, dass diese für Recall-Fehler eine deutlich größere Rolle spielen als für Präzisionsfehler, weshalb wir nur in der Kategorie A mehrere Typen von Vorverarbeitungsfehlern unterscheiden. Wir

[10] Für *false negatives* erfordert eine entsprechende Zuordnung, dass für jeden geforderten Treffer das "intendierte Suchmuster" bekannt ist. Da wir diese Information bei der Suchmustererstellung nicht erfasst haben, entfällt die Möglichkeit einer solchen Gruppierung. Die Zahlen in Tabelle 5.5 beziehen sich dementsprechend nicht auf Satz-Suchmuster-Paare sondern auf Sätze.

| A Recall-Fehler | B Präzisionsfehler |
|---|---|
| **Vorverarbeitungsbedingt** | |
| **A1** Satzgrenzenerkennung | **B1** Fehlanalyse |
| **A2** Keine Analyse | |
| **A3** Keine PreDS | |
| **A4** Parsefehler (Topo) | |
| **A5** Parsefehler (PreDS) | |
| **Suchmusterbedingt** | |
| **A6** Suchmuster zu eng | **B2** Suchmuster zu weit |
| **A7** Fehlender *frame* | |
| **A8** Fehlendes Prädikat | |
| **Sprachlich bedingt** | |
| | **B3** Ausnahme |
| | **B4** Systematische Ambiguität |
| | **B5** Kontingente Aussage |
| | **B6** Subsumtion |
| | **B7** Zweifelsfall |

Tabelle 5.6: Fehlerklassen

trennen hier nach den einzelnen in 4.2 erläuterten Stufen der Dokumentverarbeitung sowie den Verarbeitungsschritten des PreDS-Parsers. Kategorie A2 erfasst Fälle, in denen (z.B. aus Ressourcengründen) für einen zu extrahierenden Satz gar keine Analyse ermittelt werden konnte.

#### (b) A6–A8 bzw. B2: Suchmusterbedingte Fehler

Neben der Vorverarbeitung stellen die verwendeten Suchmuster eine weitere Fehlerquelle dar (A6–A8 bzw. B2). Die Gründe sind hier relativ offensichtlich. Recall-Fehler rühren von zu eng formulierten oder fehlenden Suchmustern her (wobei im Falle der dependenzbasierten Muster entweder ein Prädikat oder ein *frame* für ein bekanntes Prädikat in unserer Suchmustermenge fehlen kann). Präzisionsfehler gehen auf fehlende Einschränkungen in einem Suchmuster zu-

rück (etwa bezüglich des Aktiv / Passiv-Merkmals oder zusätzlicher Dependenzen).

#### (c) B3–B7: Sprachlich bedingte Fehler

Zudem treten aber auch noch Fehler auf, die selbst bei korrekter Vorverarbeitung und optimaler Suchmusterabdeckung verbleiben würden. Wir werden diese zusammenfassend als *sprachlich bedingte Fehler* bezeichnen. Im Falle des Recall rühren solche Fehler daher, dass für eine Definition eine Formulierung verwendet wird, zu der rein prädikatbasiert kein trennscharfes Suchmuster konstruiert werden kann. So enthält 5.5 mit *Rechnung tragen* ein Prädikat, das nur in sehr speziellen Konstellationen definitorisch verwendet wird.

(5.5) Dem Benehmenserfordernis kann aber auch durch das nachträgliche Herstellen des Benehmens Rechnung getragen werden.

(BSG 6. Senat, 1996-02-07, AZ 6 RKa 68/94, juris)

Ein Suchmuster auf Basis dieses Prädikats würde somit (unabhängig vom verwendeten *frame*) mit großer Wahrscheinlichkeit fast nur *false positives* liefern. Die Trennlinie zwischen solchen Fällen und Fällen, die ein eigenes Suchmuster rechtfertigen würden, ist anhand von Einzelbeobachtungen nur schwer zu ziehen. Wir haben deshalb für die betreffenden Recall-Fehler keine eigene Kategorie eingeführt, sondern sie der Klasse A8 zugeschlagen.

**Definiendum-bezogene Ausnahmen.** Im Falle der Präzision lassen sich nach ihrer Ursache mehr Typen sprachlich bedingter Fehler unterscheiden. Eine erste Gruppe solcher Fehler (Klasse B3 in unserem Schema) ist dadurch verursacht, dass bestimmte Begriffe zwar an Definiendum-Position in definitionstypischen Formulierungen auftreten, durch diese jedoch nicht tatsächlich inhaltlich bestimmt werden. Dies ist beispielsweise oft der Fall bei sehr allgemeinen Begriffen zum rechtlichen Status von Handlungen oder Aussagen (etwa *zulässig*, *ausgeschlossen* oder *anwendbar*). Hier ist die Angabe von Anwendungsbedingungen eher als rechtliche Norm denn als Definition aufzufassen, vgl. (5.6). Auch Aussagen mit Normverweisen an der Stelle des Definiendum haben in der Regel eher Norm- als Definitionscharakter. Sie geben wie in (5.7) oft ganz oder teilweise den Regelungsgehalt der zitierten Norm wieder.

(5.6) Die Grundsätze über die inländische Fluchtalternative sind auch dann anwendbar, wenn der Verfolgerstaat in einer Region seine Gebietsgewalt vorübergehend faktisch verloren hat.

(Oberverwaltungsgericht des Saarlandes 9. Senat, 6. Februar 2002, AZ 9 R 13/99, juris)

(5.7) Die Anwendung der §§ 127 ff. BauGB auf einen Verkehrsweg setzt nämlich voraus, dass er zum Katalog der in § 127 Abs. 1 Nrn. 1 bis 3 genannten Erschließungsanlagen gehört.

(Oberverwaltungsgericht für das Land Nordrhein-Westfalen 3. Senat, 7. September 2001, AZ 3 A 5059/98, juris)

Begriffe wie *Anwendung* oder Normzitate als Definiendum stellen für die Definitionssuche eine Art "Stoppbegriffe" dar. Fehler der Klasse B3 wären in der Regel durch entsprechende Ausnahmebedingungen auf lexikalischer Basis zu vermeiden (allerdings u.U. auf Kosten des Recall).

Die verbleibenden Typen von Präzisionsfehlern gehen auf semantisch-pragmatische Eigenschaften der fälschlicherweise extrahierten Sätze zurück. Sie hängen mehr oder weniger direkt mit theoretischen Schwierigkeiten bei der scharfen Eingrenzung des Definitionsbegriffs zusammen.

**Ambiguität einer Definitionsformulierung.** Unter B4 und B5 fassen wir Fehler, die von Ambiguitäten herrühren, bei denen Oberflächenmerkmale für das Vorliegen der einen oder anderen Lesart unter Umständen völlig fehlen. In den Fällen B4 ist das Definitionsprädikat semantisch ambig. So wird in (5.8) *bedeuten* zur Bezeichnung einer Rechtsfolge statt zur Intensionsangabe verwendet. In den Fällen B5 besteht dagegen ohne klaren semantischen Unterschied eine Ambiguität zwischen definitorischer Verwendung eines Formulierungsmusters und seiner Verwendung in einer kontingenten (d.h. nicht-definitorischen) Aussage. Beispiel (5.9) illustriert dies für die Konstruktion *A + sein + B + Relativsatz*.

(5.8) Denn die ausdrückliche Zulassung der Lagerhäuser und Lagerplätze in den Gewerbe- und Industriegebieten gemäß §§ 8 und 9 BauNVO bedeutet nicht, dass sie schon allein deshalb in allen anderen Baugebieten unzulässig sind.

(BVerwG 4. Senat, 8. November 2001, AZ 4 C 18/00, juris)

(5.9) Gaststätten sind potentielle Störungsquellen, die demjenigen, der sie eröffnet, auch zugeordnet werden müssen.

(Verwaltungsgerichtshof Baden-Württemberg 14. Senat, 20. Februar 1992, AZ 14 S 3415/88, juris)

**Nicht essentiell / abstrakt-generell.** Die Klassen B6 und B7 schließlich beinhalten Fälle, in denen weder ein Definiendum-bezogener Ausnahmefall noch

eine Ambiguität des extrahierten Satzes vorliegt, diesem jedoch der für Definitionen charakteristische essentielle und abstrakt-generelle Charakter fehlt. Die Klasse B6 zielt auf den Fall definitorisch formulierter Subsumtionen (wie in (5.10) mit dem Definitionsprädikat *darstellen*). Mit der Klasse B7 erfassen wir "pseudo-generelle" oder extrem abgeschwächte Aussagen, die entweder nur auf sehr wenige Fälle anwendbar sind oder kaum relevante Information enthalten. So wird beispielsweise in (5.11) zwar das Definitionsprädikat *umfassen* gebraucht. Für die Geltung der Aussage wird jedoch eine ausgesprochen spezielle Vorbedingung aufgestellt und außerdem im Definiens ein für die Verwendung des Definiendum *Rechtseinräumung (durch Übergabe von Fotos zum Abdruck)* kaum verwertbares Bedeutungselement angeführt.

(5.10) Soweit der Kläger sein Unterlassungsbegehren nachträglich auch auf Informationen erstreckt hat, die ihn als Dritten im Sinne des § 6 Abs. 7 StUG betreffen, stellt dies eine zulässige Klageänderung dar.

(VG Berlin 1. Kammer, 4. Juli 2001, AZ 1 A 389.00, juris)

(5.11) Werden einer Tageszeitung Fotos von freiberuflich tätigen Pressefotografen zum Abdruck im Printmedium übergeben, so umfasst diese Rechtseinräumung grundsätzlich nicht auch das Recht zur Nutzung der Fotos auf der Internet-Homepage oder in einem Internet-Archiv der Tageszeitung.

(KG Berlin 5. Zivilsenat, 24. Juli 2001, AZ 5 U 9427/99, juris)

Während bei Fehlern der Klassen B4 und B5 sprachliche Unterscheidungsmerkmale zu tatsächlichen Definitionen unter Umständen ganz fehlen, sind in den Klassen B6 und B7 in vielen Fällen zumindest Oberflächencharakteristika auszumachen, die als "weiche Kriterien" auf ein nicht-definitorisches Auftreten des jeweiligen Suchmusters hindeuten.

- Inhaltlich komplexe und umfangreiche Konditionalsätze, die wie in (5.11) andeuten, dass ein sehr spezieller Sachverhalt beschrieben wird.
- Besonders komplexe Kombinationen von Modifikatoren und / oder Negation des Definitionsprädikats, vgl. ebenfalls (5.11) (auch wenn "weniger extreme" Kombinationen von Modifikatoren und Negation in Definitionen generell durchaus häufig auftreten).
- Elemente mit deiktischem oder definit-anaphorischem Bezug, die eine Referenz auf partikulare Entitäten ("den konkreten Einzelfall") herstellen. Diese deuten oft – wie in (5.10) – darauf hin, dass in einem Satz ein Rechtsbegriff nicht definiert, sondern bereits subsumierend auf ein Element des Sachverhalts angewandt wird.

### 5.4.3 Verteilung auf die Fehlerklassen

In Tabelle 5.7 ist die Verteilung der untersuchten Extraktionsfehler auf diese Fehlerklassen angegeben. Um trotz der unterschiedlichen Stichprobengrößen einen direkten Vergleich der einzelnen Werte zu ermöglichen, haben wir neben den absoluten Zahlen auch prozentuale Anteile für die einzelnen Klassen angeführt.

Analysefehler für die sequenzbasierte Extraktion sind unter *Parsefehler (Topo)* eingetragen. Die sechs A8-Fehler der sequenzbasierten Extraktion aus dem Pilotstudien-Korpus sind darauf zurückzuführen, dass wir Instanzen in den Entwicklungsdaten bewusst nicht in ein sequenzbasiertes Suchmuster umgesetzt haben, wenn dieses außer dem Verb *sein* keine weiteren Lemmata oder Beschränkungen enthalten hätte. Sie sind daher "außer Konkurrenz" zu betrachten. Es verbleiben somit vier echte *false negatives* bei der sequenzbasierten Suche.

**Recall-Fehler.** Bei der Extraktion aus dem *Pilotstudien-Korpus* spielen Vorverarbeitungsprobleme – wie zu erwarten – die mit Abstand größte Rolle als Quelle für Recall-Fehler. Die vier echten *false negatives* der sequenzbasierten Suche sind sämtlich vorverarbeitungsbedingt. In allen Fällen wurde einem kategorieambigen Wort die falsche Wortklasse zugewiesen. Der niedrigere Recall bei der dependenzbasierten Extraktion geht ebenfalls fast vollständig auf Verarbeitungs- und Analysefehler zurück. Das größte Problem stellen Fehlanalysen des PreDS-Parsers dar, zu etwa gleichen Teilen bei der Ermittlung der topologischen Struktur und bei der PreDS-Konstruktion.

Falsche topologische Analysen rühren insbesondere von asyndetischen Satzkoordinationen und der modalen Verwendung von *sein* her. Auf PreDS-Ebene treten häufig Probleme aufgrund von Fehlern bei der heuristischen Rollenzuordnung auf. Verglichen mit Lemmatisierung und Wortklassenzuweisung ist die Analyse durch den PreDS-Parser zudem auch weniger robust: Sowohl Laufzeitfehler des Parsers (die zum Fehlen der Gesamtanalyse oder zumindest der PreDS-Struktur führen) als auch Probleme der vorgeschalteten Satzgrenzenerkennung (meist bedingt durch Satzzeichen in Abkürzungen und Normverweisen) schlagen auf das Gesamtergebnis der dependenzbasierten Extraktion durch, während diese Problemtypen auf dem sequenzbasierten Verarbeitungspfad ohne Auswirkungen bleiben.

Dagegen deuten die Recall-Fehler aus dem *Goldstandard-Korpus* bei beiden Suchmustertypen auf ein gewisses *Overfitting* auf die Entwicklungsdaten hin. Für die sequenzbasierte Extraktion sind die untersuchten Recall-Fehler vollständig auf fehlende oder zu eng formulierte Suchmuster zurückzuführen, d.h. auf das Auftreten neuer, in den Entwicklungsdaten so nicht angetroffener de-

| **Pilotstudien-Korpus** | *sequenzbasiert* | | *dependenzbasiert* | |
|---|---|---|---|---|
| **A Recall-Fehler** | | | | |
| A1 Satzgrenzenerkennung | 0 | 0% | 2 | 4% |
| A2 Keine Analyse | 0 | 0% | 1 | 2% |
| A3 Keine PreDS | n.a. | 0% | 9 | 16% |
| A4 Parsefehler (Topo) | 4 | 40% | 20 | 36% |
| A5 Parsefehler (PreDS) | n.a. | 0% | 23 | 41% |
| A6 Suchmuster zu eng | 0 | 0% | 1 | 2% |
| A7 Frame fehlt | n.a. | 0% | 0 | 0% |
| A8 Prädikat fehlt | 6 | 60% | 0 | 0% |
| **B Präzisionsfehler** | | | | |
| B1 Fehlanalyse | 9 | 4% | 25 | 11% |
| B2 Suchmuster zu weit | 155 | 66% | 27 | 12% |
| B3 Ausnahme | 5 | 2% | 41 | 18% |
| B4 Semantische Ambiguität | 6 | 3% | 15 | 7% |
| B5 Kontingente Verwendung | 25 | 11% | 57 | 25% |
| B6 Subsumtion | 14 | 6% | 20 | 9% |
| B7 Zweifelsfall | 22 | 9% | 43 | 19% |

| **Goldstandard-Korpus** | *sequenzbasiert* | | *dependenzbasiert* | |
|---|---|---|---|---|
| **A Recall-Fehler** | | | | |
| A1 Satzgrenzenerkennung | 0 | 0% | 0 | 0% |
| A2 Keine Analyse | 0 | 0% | 1 | 1% |
| A3 Keine PreDS | n.a. | 0% | 7 | 10% |
| A4 Parsefehler (Topo) | 0 | 0% | 4 | 6% |
| A5 Parsefehler (PreDS) | n.a. | 0% | 7 | 10% |
| A6 Suchmuster zu eng | 12 | 17% | 3 | 4% |
| A7 Frame fehlt | n.a. | 0% | 22 | 31% |
| A8 Prädikat fehlt | 58 | 83% | 26 | 37% |
| **B Präzisionsfehler** | | | | |
| B1 Fehlanalyse | 14 | 5% | 32 | 15% |
| B2 Suchmuster zu weit | 191 | 73% | 32 | 15% |
| B3 Ausnahme | 6 | 2% | 38 | 18% |
| B4 Semantische Ambiguität | 1 | 0% | 6 | 3% |
| B5 Kontingente Verwendung | 12 | 5% | 44 | 21% |
| B6 Subsumtion | 10 | 4% | 23 | 11% |
| B7 Zweifelsfall | 27 | 10% | 32 | 15% |

Tabelle 5.7: Fehler bei der Extraktion aus Entwicklungs- und Testdaten

finitorischer Formulierungen. Bei der dependenzbasierten Extraktion ist diese Problemklasse für etwa drei Viertel der Recall-Fehler verantwortlich.

Neben Varianten bekannter Muster (etwa Formulierungen mit *man* für Muster, die in den Entwicklungsdaten passivisch formuliert wurden) spielen hier zu etwa gleichen Teilen fehlende definitorische Valenzrahmen bei bekannten Definitionsprädikaten und fehlende Definitionsprädikate eine Rolle. Die fehlenden Prädikate sind allerdings nur zu einem Teil generell relevant (z.B. *einschließen* oder *gehen um*). Teilweise handelt es sich um eher idiosynkratische oder zumindest sehr definiendumspezifische Formulierungen (z.B. *Rechnung tragen*, *gebieten*).

Aufgrund des hohen Anteils suchmusterbedingter Recall-Fehler ist der verbleibende Stichprobenumfang für die restlichen Fehlerklassen insgesamt klein (19 Treffer). Wir gehen daher davon aus, dass die verglichen mit dem Pilotstudien-Korpus deutlich veränderte Verteilung zwischen den Klassen A3 und A4 zufallsbedingt ist.

**Präzisionsfehler.** Bei den Präzisionsfehlern spielen Vorverarbeitungsprobleme eine erheblich geringere Rolle als bei den Recall-Fehlern. Im Fall der sequenzbasierten Extraktion tritt eindeutig die mangelnde Trennschärfe der Suchmuster als Hauptfehlerquelle hervor. Die mit Abstand größten Probleme bestehen hier darin, dass mittels Beschränkungen über Lemma + POS-Abfolgen nicht eindeutig festgelegt werden kann, ob zwei Wörter im selben Teilsatz eines Satzgefüges stehen und ob ein Nomen in Kombination mit *sein* als Prädikatsnomen fungiert.[11]

Bei der dependenzbasierten Suche entfällt dagegen der größte Teil der Präzisionsfehler auf die sprachlich bedingten Fehlertypen B4–B7. Die häufigste Fehlerquelle stellen dabei Fälle dar, in denen ein Definitionsprädikat ohne Bedeutungsänderung in einer kontingenten Aussage verwendet wird (B5). Dagegen spielen semantisch ambige Definitionsprädikate (B4) für diese *false positives* die geringste Rolle.

Der Übergang auf das Goldstandard-Korpus führt im Fall der Präzisionsfehler nur zu recht geringen Veränderungen. Insbesondere treten nicht wesentlich mehr Fehler aufgrund zu weiter Suchmuster auf. Dies bestätigt, dass unsere Suchmuster (wie bei der Gesamtevaluation vermutet) keine zu weitgehenden Abstraktionen beinhalten.

[11] Diese Struktur ist ein wesentliches Element des Grundmusters einer Definition *per genus et differentiam*. Allein für mehr als die Hälfte der 21 Suchmuster mit *sein* müssten daher Prädikatsnomen korrekt erkannt werden. Da *sein* eines der häufigsten Lemmata im Korpus darstellt, rühren mehr als 80% der Fehler aus der Klasse B2 bei der sequenzbasierten Suche von dem beschriebenen Problem her.

**Bewertung.** Insgesamt hat die Fehleranalyse den bereits in 5.3.2 gewonnenen Eindruck bestätigt, dass zumindest im Fall der dependenzbasierten Suche realistisches Potential für eine Performanzverbesserung besteht. Eine Recall-Erhöhung dürfte (wie ebenfalls bereits in 5.3.2 vermutet) v.a. durch die Verbesserung der linguistischen Vorverarbeitung und weitere Suchmuster zu erreichen sein. Dagegen werden im Falle der Präzision neben den Suchmustern zusätzliche Auswahl- und Filterungsmechanismen für Treffer benötigt. Die beiden letzgenannten Punkte werden wir im nächsten Kapitel genauer untersuchen.

Zugleich deuten die Ergebnisse unserer Fehleranalyse allerdings auch darauf hin, dass ein gewisser Teil der Definitionen in Gerichtsurteilen anhand sprachlicher Merkmale prinzipiell nicht zuverlässig identifiziert werden kann. Dazu ist anzumerken, dass das *inter-annotator agreement* bei der Erstellung unseres Goldstandards eher niedrig lag (vgl. 2.4.2). Dies zeigt, dass die Aufgabe der Definitionsidentifikation auch von Menschen nur bis zu einer gewissen Grenze einheitlich gelöst wird. Hieraus resultiert eine obere Schranke für die Qualität, die von einem automatischen Extraktionsverfahren erwartet werden kann.

## 5.5 Zusammenfassung und Diskussion

Wir haben in diesem Kapitel das im CORTE-System verwendete Definitionsextraktionsverfahren und seine technische Umsetzung vorgestellt. Die Extraktionsergebnisse haben wir dann unter Verwendung verschiedener Referenzdatenbestände evaluiert. Abschließend haben wir die Ursachen für Recall- und Präzisionsfehler diagnostiziert und so Rückschlüsse auf Optimierungsmöglichkeiten ziehen können, denen wir uns im nächsten Kapitel zuwenden werden.

Die dargestellten Experimente bestätigen die Umsetzbarkeit eines Definitionsextraktionssystems auf der Grundlage einer wissensbasierten Vorgehensweise. Sie lassen außerdem klar den Mehrwert eines linguistisch informierten Ansatzes erkennen.

Auf sprachwissenschaftlich-theoretischer Ebene validieren unsere Ergebnisse die Analysen aus Kap. 2. Die hohen Präzisionswerte, die für einen Teil der Suchmuster erreicht werden, zeigen dass einige der dort identifizierten Konstruktionen sehr spezifische Definitionsindikatoren darstellen. Unsere Fehleranalyse bestätigt jedoch auch, dass eine zweite wichtige Klasse von definitorischen Formulierungen auf Prädikaten basiert, die ebenso in anderen (nichtdefinitorischen) Kontexten gebräuchlich sind. Hier können zusätzliche “weiche” Indikatoren für eine Unterscheidung zwischen Definition und kontingenter Aussage entscheidend sein. Nicht in allen Fällen sind allerdings solche Indikatoren vorhanden.

Aus praktischer Sicht ist eine Bewertung der Ergebnisqualität des CORTE-Systems natürlich nur relativ zu bestimmen Anwendungskontexten möglich. Wir gehen aufgrund der Evaluation in diesem Kapitel davon aus, dass das CORTE-System ohne große Modifikationen für Anwendungen einsetzbar ist, die eine manuelle Nachbearbeitung der Ergebnisse einschließen (z.B. die Erzeugung von Leitsatz-Kandidaten oder Schlagwort-Vorschlägen bei der Entscheidungsdokumentation). Für bestimmte vollautomatisierte Anwendungen (z.B. eine textbasierte Ontologieextraktion oder -erweiterung) besteht hingegen noch Optimierungsbedarf. Auf verschiedene Optimierungsansätze gehen wir im nächsten Kapitel ein.

Die für den praktischen Einsatz des CORTE-Systems relevante Frage der Laufzeitoptimierung konnte in diesem Kapitel nur gestreift werden. Die angesprochene Umsetzung der dependenzbasierten Definitionssuche auf der Basis eines relationalen Datenbanksystems läßt eine “offline-Nutzung” der Definitionssuche, z.B. als Vorverarbeitungsschritt bei der Indizierung einer Rechtsprechungsdatenbank, realistisch erscheinen. Bei gezielter Optimierung des Datenmodells dürfte auch der Einsatz als Echtzeit-Komponente in Reichweite rücken.

# Kapitel 6

# Ergebnisoptimierung

Viele Definitionen sind schon anhand eng eingegrenzter sprachlicher Charakteristika zu erkennen, die – wie in den letzten Kapiteln gezeigt – durch die intellektuelle Analyse einer handhabbaren Menge von Korpusbeispielen identifiziert werden können. Diese Vorgehensweise stößt jedoch zum Beispiel bei Definitionen, die nur im Kontext und anhand semantisch-pragmatischer Kriterien zu identifizieren sind, an ihre Grenzen. Dasselbe gilt für definitorische Formulierungen, die insgesamt relativ selten auftreten oder nur in Einzelfällen eine Definition anzeigen. Auch häufigere und stärker standardisierte Formulierungsmuster sind zudem unterschiedlich spezifisch als Definitionsanzeiger.

Solche Probleme wirken sich auch auf die Performanz des im vorigen Kapitel beschriebenen Definitionsextraktionsverfahrens aus. Der Kernbestand an Definitionsmustern, den wir aus der Analyse unseres Entwicklungskorpus von vierzig Entscheidungstexten gewinnen konnten, reicht zwar bereits aus, um relativ zuverlässig eine interessante Zahl von Definitionen zu identifizieren. Der Recall wird jedoch wesentlich durch das Fehlen von Suchmustern auf der Basis seltenerer, beispielsweise nur mit bestimmten Definienda assoziierter Definitionsprädikate eingeschränkt, und aus dem Fehlen von Zusatzkriterien in den vorhandenen Suchmustern resultieren Präzisionsfehler. Für einen praktischen Einsatz, etwa zur Unterstützung von Benutzern einer Rechtsprechungsdatenbank, dürfte die Leistungsfähigkeit des Systems in der in Kap. 5 beschriebenen Form deshalb nur bedingt ausreichen.

In diesem Kapitel untersuchen wir verschiedene Möglichkeiten zur Verbesserung der Ergebnisqualität unseres Definitionsextraktionssystems. Zunächst diskutieren wir in 6.1 die Resultate einiger naheliegender, jedoch eher punktueller Verbesserungsmaßnahmen:

- Optimierung der Präzision durch Auswahl von besonders präzisen Suchmustern
- Optimierung der Präzision durch Stoppwort-basierte Filterung der Treffer

- Optimierung des Recall durch systematische Identifikation und Nachpflegen fehlender Suchmuster

Diese Maßnahmen sind zwar technisch einfach zu realisieren, jedoch durchgängig sehr wissensintensiv. Ihre Umsetzung erfordert daher im Vorfeld erheblichen Analyseaufwand und ist somit nur in begrenztem Umfang praktikabel. Zudem führten sie zumindest in unserem Fall nicht zu Verbesserungen, die den betriebenen Aufwand gerechtfertigt erscheinen ließen.

Im zweiten Teil des Kapitels diskutieren wir dann mehrere automatische, datengestützte Optimierungsansätze. Den besprochenen Verfahren ist eine vergleichsweise höhere technische Komplexität gemeinsam. Diese wird jedoch aufgewogen durch das Potential, mit einem geringen manuellen Aufwand auf einer breiten Datengrundlage eine hohe Qualitätssteigerung zu erzielen. In 6.2 beschreiben wir den Einsatz von Klassifikations- und Ranking-Verfahren, die sich auf eine Vielzahl von Kriterien stützen, zur Identifikation der *true positives* unter den Treffern unseres Extraktionssystems. In 6.3 befassen wir uns mit einem Bootstrappingverfahren zur automatischen Identifikation zusätzlicher Suchmuster. Durch die kombinierte Anwendung beider Verfahren in 6.3.2 können wir eine erhebliche Ergebnisverbesserung erzielen.

## 6.1 Punktuelle Verbesserungsmaßnahmen

Wir diskutieren in diesem Abschnitt Experimente zur Optimierung unserer Definitionssuche, die direkt an mehreren im vorigen Kapitel diagnostizierten Problemen ansetzen. Unsere Fehleranalyse dort hat ergeben, dass (neben Mängeln der linguistischen Vorverarbeitung) Defizite der aus der Datenanalyse in Kap. 2 hergeleiteten Suchmuster die Ergebnisqualität begrenzen.

Für Präzisionsprobleme, d.h. das Auftreten von *false positives* in den Suchergebnissen, haben wir zwei wichtige Ursachen ausgemacht, die wir im Folgenden gezielt angehen werden:

- Von den einzelnen Suchmustern werden sehr unterschiedliche Präzisionswerte erzielt. Ein Teil der Suchmuster ist aufgrund der zu geringen Spezifizität für eine akkurate Identifikation von Definitionen schlicht ungeeignet. Wir erproben daher hier zunächst Verfahren zur Auswahl einer Untermenge aus dem im vorigen Kapitel verwendeten Suchmusterbestand, die möglichst wenige dieser ungeeigneten Muster enthält.

- In vielen Fällen können *false positives* an bestimmten typischen Ausdrücken erkannt werden (z.B. deuten die Wörter *beklagt* oder *Kläger* oft

auf eine Aussage hin, die nicht allgemeingültig, sondern auf den konkreten Fall bezogen ist). Durch unsere Suchmuster, die ja außer den Definitionsprädikaten keine Inhaltswörter enthalten, sind solche Ausdrücke nicht erfasst. Wir erproben einen Filtermechanismus auf der Grundlage von Stoppwortlisten zur Entfernung solcher Fälle aus den Treffermengen unseres Extraktionssystems.

Die Untersuchung von *false negatives* hingegen wies fehlende Extraktionsmuster auch für einige gängige Definitionsprädikate als eine Hauptfehlerquelle aus. Zur Behebung dieses Problems erproben wir in 6.1.2 die Möglichkeit, orientiert an Frequenzdaten aus dem CORTE-Großkorpus eine begrenzte Anzahl besonders "aussichtsreicher" zusätzlicher Muster manuell zu spezifizieren.

### 6.1.1 Präzisionsoptimierung

#### (a) Musterauswahl

Wie im vorigen Kapitel angesprochen unterscheiden sich die einzelnen von uns verwendeten Suchmuster deutlich im Hinblick auf ihre Präzision. Sie decken den gesamten Bereich zwischen den Präzisionswerten 0 und 1 ab. Da die Trefferzahl dabei ohne direkte Abhängigkeit von der Präzision ebenfalls stark zwischen den Suchmustern variiert, erzielen verschiedene Untermengen unseres Suchmusterbestands deutlich unterschiedliche f-Scores. Durch eine geeignete Musterauswahl kann potentiell ein höherer f-Score erzielt werden als durch die Gesamtheit der bisher genutzten Suchmuster.

Ein exhaustiver Vergleich aller Untermengen ist aufgrund der Größe des entsprechenden Suchraums kein gangbarer Weg zur Identifikation des optimalen Suchmustersatzes. Die Suchmusterauswahl kann nur mittels eines heuristischen Verfahrens getroffen werden. Wir haben in einer Experimentreihe die Konstruktion einer solchen Auswahl durch ein vereinfachtes Suchverfahren auf der Basis verschiedener lokaler Evaluationsfunktionen getestet. Ausgehend von einer leeren Suchmustermenge wurde iterativ jeweils das nach der Evaluationsfunktion bestbewertete Suchmuster zum bereits gebildeten Bestand hinzugenommen. In jeder Iteration wurden Präzision, Recall und f-Score der bis dahin erzeugten Suchmustermenge ermittelt und schließlich aus den Mengen aller Iterationen die nach f-Score optimale ausgewählt. Die verwendeten Evaluationsfunktionen sind im Folgenden zusammengestellt:[1]

*Präzision:* Jedes Muster wird nach der von ihm allein erzielten Präzision bewertet

[1] Für die Berechnung der *Information Gain*-, $\chi^2$- und CFS-Metriken haben wir Komponenten aus der WEKA-Toolsuite (Hall u. a. (2009)) eingesetzt.

*Recall:* Jedes Muster wird nach dem von ihm allein erzielten Recall bewertet

*f-Score:* Jedes Muster wird nach dem von ihm und den bisher ausgewählten gemeinsam erzielten f-Score bewertet.

*Information Gain:* Muster werden als binäre Attribute der Sätze im Korpus aufgefasst (jeder Satz ist entweder unter den Treffern eines Musters oder nicht) und jedes Muster wird nach dem Zuwachs an Information bewertet, den das Vorliegen des entsprechenden Attributs bezüglich der Klassifizierung [±Definition] bedeutet.

**$\chi^2$:** Muster werden als binäre Attribute aufgefasst, und jedes Muster wird nach seinem Assoziationsgrad mit der Eigenschaft [±Definition] (gemessen durch Pearson's $\chi^2$-Koeffizienten) bewertet.

*CFS (Correlated Feature Subset):* Es wird jeweils dasjenige Muster hinzugenommen, das am stärksten mit der Eigenschaft [±Definition] korreliert ist und dabei die geringste Korrelation mit den bereits ausgewählten Mustern aufweist.

Die ersten drei Funktionen berechnen mit p, r und f die zu optimierenden Größen selber jeweils auf der Ebene einzelner Suchmuster. Dagegen handelt es sich bei der zweiten Gruppe um Funktionen, die den statistischen Zusammenhang zwischen den Suchmustern und der Zieleigenschaft [±Definition] anhand gängiger Assoziationsmaße modellieren.

In Tabelle 6.1 sind die mit den verschiedenen Evaluationsfunktionen erzielten Ergebnisse gegenübergestellt. Hier und im Folgenden haben wir Präzision und Recall nach der Methode *Erweitert 2* (siehe 5.3.1) berechnet. Als Baseline sind noch einmal die Scores für die Extraktion mit allen Mustern angeführt (vgl. Tabelle 5.2). Jedes Auswahlverfahren wurde dabei in zwei idealisierten und einem realistischen Szenario ausgewertet: Für die Ergebnisse in Spalte 1 (*Pilot / Pilot*) wurde unser Pilotstudien-Korpus sowohl für die Musterauswahl (also beispielsweise zur Ermittlung der Präzisions- bzw. Recall-Werte für die ersten beiden beschriebenen Methoden) als auch für die Extraktion genutzt (vgl. 5.3.2), für Spalte 2 (*Gold / Gold*) wurden für beide Aufgaben das Goldstandard-Korpus verwendet. In Spalte 3 (*Gold / Pilot*) wurden die Pilotstudien-Daten für die Musterauswahl zu Grunde gelegt, die Goldstandarddaten dienten zur Evaluation.

Die Ergebnisse in den Szenarios *Pilot / Pilot* und *Gold / Gold* blenden also den Faktor *Variabilität der Datengrundlage zwischen Musterbewertung und Extraktion* aus. Die Differenzen zwischen den Szenarios *Gold / Gold* und *Gold / Entwicklung* rühren von solchen Abweichungen zwischen Pilotstudien-

und Goldstandarddaten her. Die ersten beiden Szenarios können somit als eine ungefähre obere Schranke der mit dem jeweiligen Verfahren möglichen Verbesserung bei optimalen Trainingsdaten betrachtet werden. Die Ergebnisse in Spalte 3 lassen dagegen Rückschlüsse auf die Leistungsfähigkeit des Verfahrens im allgemeinen Fall zu.

| | **Test- / Trainingsdaten** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | *Pilot / Pilot* | | | *Gold / Gold* | | | *Gold / Pilot* | | |
| | *p* | *r* | *f* | *p* | *r* | *f* | *p* | *r* | *f* |
| **Baseline** | | | | | | | | | |
| *(alle Muster)* | *0,28* | *0,69* | *0,40* | *0,19* | *0,45* | *0,27* | *0,19* | *0,45* | *0,27* |
| **Auswahlmethode** | | | | | | | | | |
| *Präzision* | 0,77 | *0,46* | *0,57* | 0,40 | *0,29* | *0,34* | 0,30 (*0,28*)[2] | *0,13* *0,32* | *0,18* *0,30* |
| *Recall* | 0,31 | *0,63* | *0,42* | 0,26 | *0,37* | *0,31* | 0,20 | *0,36* | *0,26* |
| *Information Gain* | 0,46 | *0,47* | *0,46* | 0,26 | *0,36* | *0,31* | 0,20 | *0,42* | *0,27* |
| $\chi^2$ | 0,54 | *0,50* | *0,52* | 0,29 | *0,33* | *0,31* | 0,20 | *0,42* | *0,27* |
| *f-Score* | 0,35 | *0,62* | *0,45* | 0,26 | *0,37* | *0,31* | 0,21 | *0,35* | *0,26* |
| *CFS* | 0,43 | *0,60* | *0,50* | 0,28 | *0,35* | *0,31* | 0,24 | *0,27* | *0,26* |

Tabelle 6.1: Ergebnisse verschiedener Methoden zur Suchmusterauswahl

In den optimalen Szenarios 1 und 2, in denen zur Musterauswahl und zur Evaluation ein identische Datengrundlage genutzt wurde, erbringen alle Auswahlmethoden Verbesserungen der Präzision. Diese gleichen die Verminderungen des Recall (durch die weggelassenen Suchmuster) so weit aus, dass der f-Score der optimierten Suchmustermengen den aller Suchmuster übertrifft. Der durch Auswahl der jeweils einzeln präzisesten Suchmuster erzielte f-Score stellt in diesen Szenarios den höchsten allein auf der Basis einer Musterauswahl erzielbaren f-Score dar.[3]

In dem realistischen Szenario 3, in dem die Testdaten nicht zur Musterauswahl, sondern nur zur Evaluation herangezogen wurden, führt hingegen zunächst keine Auswahlmethode zu einer Verbesserung des f-Score. Zwar kann durch eine Musterauswahl nach Präzision, f-Score sowie der CFS-Methode jeweils eine gewisse Verbesserung der Präzision erzielt werden, jedoch gleicht dieser Zugewinn in keinem der Fälle den mit der Selektion einhergehenden

[2] Präzisionsschätzungen auf der Grundlage des CORTE-Großkorpus

[3] Da hier (im Gegensatz zum Szenario *Gold / Pilot*) Test- und Entwicklungsdaten identisch sind, handelt es sich bei den zur Musterauswahl genutzten Präzisionswerten um exakte Werte und nicht um Schätzungen. Das erzeugte Muster-Ranking ist somit optimal.

Recall-Verlust aus. Insbesondere die (in den anderen Szenarios optimale) Musterauswahl nach Präzision führt zu einem starken Einbruch des f-Score. Der Grund hierfür liegt offenbar darin, dass durch die Präzisionsschätzungen anhand des relativ kleinen Pilotstudien-Korpus für diese Datengrundlage spezifische Muster zu stark präferiert werden. Die Einbeziehung der annotierten Treffermengen aus dem CORTE-Großkorpus (vgl. Tabelle 5.2 in 5.3.2) bei der Präzisionsschätzung führt – wie in Tabelle 6.1 zusätzlich angegeben – zu einer Musterauswahl, mit der auch im Szenario *Gold / Pilot* eine Präzisionsverbesserung erzielt werden kann, die (in kleinem Umfang) auf den f-Score der selektieren Suchmustermenge durchschlägt.

#### (b) Filterung

Eine weitere wesentliche Ursache für Präzisionsprobleme, die unsere Fehleranalyse im letzten Kapitel aufgezeigt hat, stellen "Pseudo-Definitionen" dar, d.h. Sätze, die zwar einem typischerweise definitorischen Formulierungsmuster folgen, bei denen es sich aber z.B. aufgrund des in der Definiendum-Position auftretenden Begriffs oder aufgrund eines zu starken Einzelfallbezugs nicht um Definitionen handelt. Häufig ist diese bei Sätzen der Fall, in denen ein Definitionsprädikat verwendet wird, weil es kollokativ mit einem Satzteil assoziiert ist (z.B. *Mangel* und *bestehen*). Außerdem können semantische Eigenschaften des Begriffs in Definiendum-Position eine Definition im klassischen Sinne ausschließen. Neben einigen sehr allgemeinen oder abstrakten Begriffen (z.B. *Wesen* oder *Einfluss*), institutionellen Rechtsbegriffen und Bezeichnungen für Teile der Rechtsordnung (Normen, Gerichte) stellen auch Nominalphrasen mit bestimmten Adjektiven (v.a. zum Ausdruck sehr allgemeiner rechtlicher Qualifikationen, z.B. *zulässig* oder *pflichtwidrig*) oft solche "Nicht-Definienda" dar. Schließlich wird auch die Subsumtion, also die rechtliche Einordnung des Einzelfalls, oft durch normalerweise definitorische Formulierungen zum Ausdruck gebracht. Der Bezug auf den konkreten Einzelfall bringt es mit sich, dass in solchen Sätzen fast immer deiktische Adverbien auftreten oder das "Subsumendum" Definita, Demonstrativpronomen bzw. Adjektive mit demonstrativer Funktion enthält.

"Pseudo-Definitionen" lassen sich also in vielen Fällen am Auftreten bestimmter Wörter erkennen. Solche *false positives* können prinzipiell durch eine einfache lexikalische Filterung aus den Ergebnissen der Definitionssuche entfernt werden, wenn entsprechende Stoppwortlisten verfügbar sind. Um das Potential eines solchen Filterungsmechanismus für die Verbesserung der Präzision unserer Extraktionsergebnisse zu ermitteln, haben wir drei Stoppwortlisten zusammengestellt:

1. *Initiale Stopp-Definienda:* Für diese Stoppwortliste wurden wissensbasiert 75 Substantive und 90 Adjektive zusammengestellt (initiale Stoppwörter), die den o.g. Kriterien entsprechen (z.B. *Wesen*, *Normadressat*, *allgemein*).

2. *Erweiterte Stopp-Definienda:* Beide Listen mit Stopp-Definienda wurden dann mittels einer Ähnlichkeitssuche auf einem Teil der geparsten Urteilstexte aus dem CORTE-Großkorpus um jeweils 450 Einträge erweitert.[4] Intuitiv erschienen viele der Ergänzungen (z.B. *Verfahrensablauf* oder *Vorgehen*) erfolgversprechend als Stoppwörter, in manchen Fällen handelte es sich jedoch auch um Begriffe, für die ohne weiteres Definitionen auftreten könnten (z.B. *Gefahrerforschungsmaßnahme* oder *Linienbestimmung*).

3. *Subsumtions-Signale:* Als solche wurden aufgrund ihrer referentiellen Funktion der bestimmte Artikel, die Demonstrativpronomen und das Pronominaladjektiv *solch* gewertet, zudem Auftreten der Wörter *vorliegend*, *hier*, *Kläger* und *beklagt*, die inhaltlich auf die Thematisierung eines konkreten Einzelfalls hindeuten.

In Tabelle 6.2 sind die nach der Filterung mit den Stoppwortlisten (in verschiedenen Kombinationen) für die Extraktion aus dem Goldstandard-Korpus erreichten Präzisions-, Recall- und f-Scores angegeben. Neben den Werten für die Extraktion mit allen Mustern ist auch das Ergebnis für die Filterung der Extraktionsresultate der besten Suchmusterauswahl aus Abschnitt (a) angegeben (präzisionbasierte Auswahl mit den annotierten Treffern aus dem CORTE-Großkorpus als Grundlage der Präzisionsschätzungen). Baseline sind die Scores für die ungefilterten Ergebnismengen (vgl. Tabelle 5.2 für alle Suchmuster und Tabelle 6.1 für die Suchmusterauswahl).

Es ergeben sich durch die Filterung teilweise erheblich erhöhte Präzisionswerte. Die verwendeten Wortlisten sind jedoch offenbar nicht sehr trennscharf. Durch die unterdrückten korrekten Treffer sinkt in allen getesteten Szenarios der Recall so stark, dass eine Verschlechterung des f-Score gegenüber der ungefilterten Extraktion zu verzeichnen ist. Am deutlichsten ist dieser Effekt im Falle der Anwendung aller Filter auf die Extraktionsergebnisse der Suchmustermenge zu erkennen, die nach dem Präzisionsranking als optimal bewertet

[4] Hierfür wurden Kontextvektoren auf der Grundlage aller direkten Dependenzen für die manuell zusammengestellten Stoppwörter sowie sämtliche Substantive und Adjektive mit mehr als zehn Auftreten erzeugt. Auf dieser Basis wurden dann die nach dem Kosinus-Maß nächsten Nachbarn der bekannten Stoppwörter gesucht (unter Verwendung der in Pado und Lapata (2007) beschriebenen Werkzeuge). Für Erläuterungen zur Nutzung von Ähnlichkeitsmaßen und Vektorraum-Modellen bei der Informationssuche vgl. 3.1.1.

| | **Alle Suchmuster** | | | **Optimale Menge** (Absatz (a), *präzisionsbasierte Auswahl)* | | |
|---|---|---|---|---|---|---|
| | *p* | *r* | *f* | *p* | *r* | *f* |
| **Baseline** *(keine Filterung)* | *0,19* | *0,45* | *0,27* | *0,28* | *0,32* | *0,30* |
| **Filterungsmethode** | | | | | | |
| *Initiale Stopp-Definienda* | 0,22 | 0,34 | 0,27 | 0,28 | 0,28 | 0,28 |
| *Erweiterte Stopp-Definienda* | 0,22 | 0,24 | 0,23 | 0,28 | 0,21 | 0,24 |
| *Subsumtionssignale* | 0,25 | 0,26 | 0,26 | 0,37 | 0,18 | 0,24 |
| *Initiale Stopp-Definienda + Subsumtionssignale* | 0,30 | 0,20 | 0,24 | 0,37 | 0,16 | 0,22 |
| *Erweiterte Stopp-Definienda + Subsumtionssignale* | 0,33 | 0,14 | 0,20 | 0,40 | 0,12 | 0,18 |

Tabelle 6.2: Ergebnisse der Extraktion aus dem Goldstandard-Korpus nach Filterung

wurde (Spalte 2). Hier ergibt sich eine Präzisionssteigerung von mehr als einem Drittel (von 0,28 auf 0,4). Durch den verminderten Recall sinkt der f-Score jedoch zugleich um etwa ein Drittel (von 0,3 auf 0,18).

### 6.1.2 Recall-Verbesserung durch Erweiterung der Suchmustermenge

Ein erheblicher Teil der *false negatives* (d.h. der nicht gefunden Definitionen) bei der Extraktion aus unserem Goldstandard-Korpus geht gemäß unserer Fehleranalyse im vorigen Kapitel auf das Fehlen von Suchmustern zurück. Ursache hierfür ist der relativ geringe Umfang des Entwicklungskorpus (auf dem durch die genutzten Muster naturgemäß eine erheblich besserer Recall erzielt wird). Zu etwa gleichen Teilen fehlen in unserem Suchmusterbestand Definitionsprädikate (die also in den Entwicklungsdaten überhaupt nicht in definitorischer Verwendung angetroffen wurden) und definitorische Valenzrahmen bei bekannten Definitionsprädikaten.

Den naheliegendsten Ansatz zur Verbesserung des Recall stellt also die Spezifikation zusätzlicher Suchmuster dar. Einer Erweiterung des Suchmusterbe-

standes nach der bisher verwendeten kombinierten wissensbasierten und korpusgestützten Methodologie (Korpusannotation und detaillierte Analyse) sind aufgrund des erforderlichen Arbeitsaufwands enge Grenzen gesetzt. Uns standen im Rahmen der hier beschriebenen Arbeiten keine Ressourcen für umfangreichere Annotationsarbeiten zur Verfügung. Als alternative Methode haben wir daher einen rein wissensbasierten, jedoch durch Frequenzdaten aus dem CORTE-Großkorpus (vgl. 4.3.1) gesteuerten Ansatz für die Spezifikation zusätzlicher Suchmuster erprobt. Aus den insgesamt 449 Verben mit 1000 und mehr Vorkommen in diesem Korpus haben wir sämtliche nach ihrer Semantik introspektiv als möglicherweise definitorisch eingeschätzten ausgewählt. Für die uns am relevantesten erscheinenden Valenzrahmen haben wir Suchmuster erstellt. Hierbei ergaben sich insgesamt 44 neue Suchmuster (für 44 Verben), die wir der initialen, korpusgestützt entwickelten Suchmustermenge hinzugefügt haben.

In Tabelle 6.3 sind die Ergebnisse für die Anwendung dieser neuen Suchmuster auf das Goldstandard-Korpus sowie für verschiedene Kombinationen mit bereits getesteten Suchmustermengen zusammengestellt.

| *Suchmuster* | *p* | *r* | *f* |
|---|---|---|---|
| **Neu** | | | |
| *Gesamt* | 0,14 | 0,03 | 0,05 |
| **Neu + bekannt** | | | |
| *Gesamt* | 0,19 | 0,44 | 0,27 |
| *Optimale Untermenge* | 0,25 | 0,35 | 0,29 |
| *Optimale Untermenge (gefiltert)* | 0,26 | 0,16 | 0,20 |

Tabelle 6.3: Extraktionsergebnisse mit manuell spezifizierten zusätzlichen Suchmustern

Aufgrund der zusätzlich eingeführten *false positives* ergibt sich in den kombinierten Szenarios durchgängig eine Verschlechterung des f-Score. In dem von uns betrachteten konkreten Testfall haben die zusätzlich spezifizierten Suchmuster somit keine Verbesserung der Ergebnisqualität erbracht. Wie der sehr geringe Recall-Wert von 0,03 für die Extraktion mit den neuen Suchmustern alleine erkennen lässt, sind unter den zusätzlichen Definitionsverben offenbar nur wenige, die überhaupt in Definitionen in unserem Goldstandard-Korpus auftreten. In keinem der kombinierten Szenarios führen die hinzugenommenen

Suchmuster zudem tatsächlich zu einem Anstieg des Recall um diese vollen drei Prozentpunkte.[5]

### 6.1.3 Analyse und Bewertung

Die bisher betrachteten Ansätze zur Optimierung der Definitionssuche haben nur zu geringem Erfolg geführt. Die Präzision der Suche lässt sich zwar durch eine geeignete Suchmusterauswahl erheblich erhöhen. Ein weiterer Präzisionszuwachs wird durch einen Stoppwort-basierten Filtermechanismus erzielt. Beide Schritte führen jedoch auch zu einem Verlust korrekter Suchergebnisse. Dieser ist so groß, dass im ersten Fall trotz des Präzisionsgewinns nur ein geringfügiger Zuwachs im f-Score gegenüber der unpräziseren Suche mit allen Mustern zu verzeichnen ist. Im zweiten Fall überwiegt der Recall-Verlust den Präzisionsgewinn im f-Score sogar.

Hierfür lassen sich mehrere Ursachen ausmachen. Zunächst einmal ist zu vermuten, dass die betrachteten Informationsquellen (Abschätzungen zur Zuverlässigkeit der einzelnen Suchmuster und sehr grobe lexikalische Merkmale einzelner Treffer) für eine deutlichere Ergebnisverbesserung zu undifferenziert und nicht ergiebig genug sind. So bleiben "weichere" Definitionsmerkmale bei der Trefferselektion nach den bisher angesprochenen Auswahlverfahren (wie schon im eigentlichen Extraktionsprozess) außer Betracht. Im Falle der Suchmusterauswahl bezieht sich die genutzte Information zudem gar nicht auf einzelne Treffer, sondern auf ganze Treffermengen. Eine weitere Schwierigkeit stellt die der Arbeitsweise eines Filters inhärente "alles oder nichts"-Entscheidung dar (Beibehaltung oder vollständige Entfernung eines Treffers bzw. sogar der gesamten Treffermenge eines Suchmusters). Eine besserer Ausgleich zwischen Präzisionsverbesserung und Recall dürfte durch Verfahren zu erzielen sein, die für die Trefferauswahl eine differenzierte Entscheidung auf der Basis einer breiten Kombination von Merkmalen und Informationsquellen treffen können. Dies ist vor allem mit induktiven Verfahren möglich, die in der Lage sind, entsprechend flexible Entscheidungsregeln aus großen Datenmengen automatisch zu gewinnen.

Hinsichtlich des Recall der Definitionssuche konnten durch die manuelle Erweiterung des Suchmusterbestandes nur marginale Verbesserungen erreicht werden. Diese gingen zudem in allen betrachteten Szenarios mit einem Präzisionsverlust einher, der insgesamt zu Verlusten im f-Score führte.

[5] Dies ist auf die Tatsache zurückzuführen, dass wir bei der Recall-Berechnung ganze Definitionen als Einheit zu Grunde legen (Szenario Erweitert 2 in 5.3.1). Treffer auf der Basis der neuen Suchmuster können daher nicht nur in bisher noch gar nicht extrahierten Definitionen liegen, sondern auch in weiteren Sätzen bereits bekannter Definitionen (bei komplexen Sätzen auch in anderen Gliedsätzen alter Treffer)

Um aus diesem Resultat eine Aussage über die generelle Brauchbarkeit des erprobten frequenzorientierten Ansatzes der Mustergewinnung abzuleiten, wäre zwar noch im einzelnen zu überprüfen, welche der neuen Suchmuster zu wie vielen weiteren Treffern geführt haben. Zudem wäre eine Überprüfung der Stabilität des von uns ermittelten negativen Ergebnisses durch die Evaluation anhand eines größeren Testkorpus notwendig. Die Ergebnisse in Tabelle 6.2 unterstützen allerdings die bereits in Kap. 5 diskutierte Beobachtung, dass der Recall-Verlust beim Übergang vom Entwicklungs- auf das Testkorpus auch auf Auftreten seltener, mehr oder weniger idiosynkratischer Definitionsformulierungen zurückzuführen ist und somit nicht allein durch die Ermittlung weiterer häufiger Definitionsmuster behoben werden kann. Nicht zuletzt aus diesem Grund ist auf der Basis einer intellektuellen Analyse von Korpusbeobachtungen wohl generell nur ein begrenzter Grad an Vollständigkeit bei der Suche erreichbar. Um auch im Hinblick auf seltenere definitorische Formulierungen eine bessere Abdeckung zu erzielen, müssen für die Mustergewinnung Korpora von solchem Umfang zu Grunde gelegt werden, dass auch hier automatisierte Verfahren benötigt werden.

## 6.2 Optimierung der Präzision durch Klassifikation und Rankings

Wir untersuchen im Rest des Kapitels nun Möglichkeiten, Ergebnisverbesserungen für die Definitionsextraktion gestützt auf datengetriebene maschinelle Lernverfahren zu erreichen. In diesem Abschnitt diskutieren wir zunächst die Ergebnisse mehrerer Experimente, bei denen wir den Einsatz automatischer Klassifikationsverfahren zur Auswahl optimierter Treffermengen unseres Definitionsextraktionssystems erprobt haben. Auf dieser Grundlage lassen sich durch die flexible Auswertung einer Vielzahl von Merkmalen der einzelnen Treffer Ergebnismengen von deutlich höherer Qualität erzeugen als durch die oben diskutierten Verfahren. Auch die höchsten erzielten Präzisionszugewinne führen nicht zu vergleichbar dramatischen Recall-Verlusten, wie sie in 6.1 zu verzeichnen waren.

Anschließend stellen wir ein Ranking-Modell vor, mit dem sich – auf der Grundlage der auch für die Klassifikation verwendeten Merkmale – Treffermengen so sortieren lassen, dass wahlweise kleinere, sehr präzise, oder größere, weniger präzise Treffermengen abgefragt werden können. Der f-Score der nach diesem Ranking erzeugten optimalen Treffermenge liegt nahe bei dem optimalen durch Klassifikation erzielbaren. Durch Einbeziehung der Voten verschiedener Klassifizierer als Informationsquelle für das Treffer-Ranking ließen

sich in unseren Experimenten gegenüber den Klassifikationsergebnissen noch geringfügig verbesserte f-Scores erzielen.

### 6.2.1 Komponenten und Informationsfluss bei der automatischen Klassifikation

In Abb. 6.1 sind Komponenten und Informationsfluss bei einer typischen automatischen Klassifikationsaufgabe dargestellt, wie sie auch dem im Folgenden verwendeten Ansatz zu Grunde liegen. Die Instanziierungen der Komponenten in unserem Fall sind in der Graphik in Klammern angegeben. In der Trainingsphase wird auf der Grundlage einer Menge von Daten mit bereits bekannter Klassenzuordnung (in unserem Falle also Sätzen, von denen bereits bekannt ist, ob es sich um Definitionen handelt oder nicht) durch ein Lernverfahren ein Modell erzeugt. Mittels dieses Modells können dann in der Klassifikationsphase neue Daten, d.h. solche mit unbekannter Klassenzugehörigkeit, den in den Trainingsdaten vertretenen Klassen zugeordnet werden (also z.B. Sätze als $[+\text{definitorisch}]$ oder $[-\text{definitorisch}]$ eingeordnet werden).

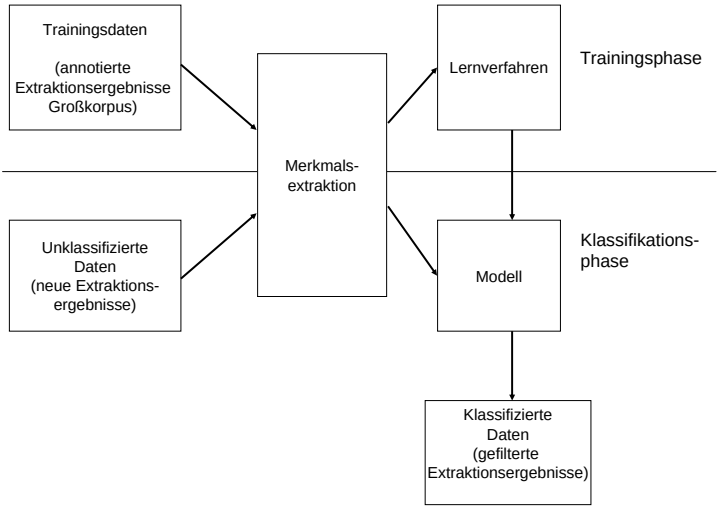



Abbildung 6.1: Komponenten und Informationsfluss bei der automatischen Klassifikation

Trainings- und neue Daten müssen dabei kompatibel und in einer für das jeweilige Lernverfahren geeigneten Weise repräsentiert werden. Die meisten Lernverfahren basieren auf Repräsentationen in Form von Merkmalsvektoren, einige erlauben dabei nur bestimmte Typen von Merkmalen. Durch die Wahl des Lernverfahrens ist jedoch nicht determiniert, welche Information in Merkmalen repräsentiert und wie diese dabei kodiert wird.

Die Wahl des Lernverfahrens und der verwendeten Merkmale sind die wesentlichen Einflussfaktoren für die Qualität der Ergebnisse einer automatischen Klassifikationsaufgabe. Einen dritten Einflussfaktor stellen schließlich die verwendeten Trainingsdaten selber dar. Für die beiden erstgenannten Entscheidungen lassen sich kaum allgemeingültige Regeln angeben. Sie können in den meisten Fällen nur unter Zuhilfenahme eines beträchtlichen Umfangs von *trial-and-error* Experimenten getroffen werden. Steigerungen in Umfang und Qualität der Trainingsdaten schlagen sich hingegen in der Regel zunächst deutlich in besseren Ergebnissen nieder, der Effekt ebbt jedoch ab einem gewissen (von Anzahl und Verteilung der genutzten Merkmale sowie dem gewählten Lernverfahren abhängigen) Punkt ab.

Als Trainingsdaten für die im Folgenden beschriebenen Klassifikationsexperimente diente uns eine Untermenge von etwa 80% (igs. 4768 Instanzen) der ca. 6000 ausgewerteten dependenzbasierten Treffer aus dem CORTE-Großkorpus.[6] Siehe 5.3.2 zur Stabilität der Annotation und somit zur Qualität der Klassifikation der Trainingsdaten.

Wir gehen nun zunächst näher auf die Merkmale ein, die wir zur Repräsentation der Trainingsdaten und der zu bewertenden neuen Treffer in der untersuchten Klassifikationsaufgabe verwendet haben. Dann vergleichen wir die Ergebnisse, die unter Nutzung verschiedener Typen von Klassifizierern erzielt wurden und bewerten auf dieser Basis das Potential von Klassifikationsverfahren zur Ergebnisverbesserung in unserem Definitionsextraktionsverfahren.

### 6.2.2 Merkmalsextraktion

In Tabelle 6.4 sind die Merkmale zusammengefasst, die wir in unseren Klassifikationsexperimenten zur Repräsentation der Treffer unserer Definitionssuche verwenden. Eine ausführlichere Aufstellung mit zusätzlichen Erläuterungen findet sich in Anhang E. Bei den Merkmalen handelt es sich zum größten Teil um annähernde Repräsentationen linguistischer Eigenschaften, die wir bereits an verschiedenen Stellen neben den “harten” Kriterien *Definitionsprädikat* und entsprechender *Valenzrahmen* als “weiche” Definitionsindikatoren in Betracht

[6] Wir haben mit den Klassifikationsexperimenten zu einem recht frühen Zeitpunkt unserer Arbeiten begonnnen. Es war deshalb erst ein Teil der Trefferannotation für das CORTE-Korpus abgeschlossen.

gezogen haben (vgl. etwa Kap. 2). Es handelt sich also um Eigenschaften, die weder hinreichend noch notwendig für die Identifikation (oder den Ausschluss) eines Satzes als Definition sind. Sie deuten jedoch typischerweise oder zumindest in bestimmten Konstellationen auf das Vorliegen einer Definition hin (bzw. lassen es besonders unwahrscheinlich erscheinen).

*Lexikalisch*

- Stoppwörter
- Boostwörter
- Subsumtionssignale
- Negation
- Modifikator
- Lexikalische und morphologische Ähnlichkeit der Definitionsbestandteile

*Referentiell*

- Definites Definiendum
- Definiter Genusbegriff
- Anaphorisches Definiendum
- Oberbegriff des Definiendum im Definiens
- Synonym des Definiendum im Definiens

*Strukturell*

- Definiendum vor Definiens
- Definition kein Hauptsatz

*Dokumentstruktur*

- Relative Dokumentposition
- Treffer im Vorkontext
- Treffer im Nachkontext

*Sonstige*

- Satzlänge
- Rechtsprechungszitat
- TF / IDF[7]-Wert des Definiendum
- Präzisionsschätzung

Tabelle 6.4: Für Trefferklassifikation und -ranking verwendete Merkmale

Die Merkmale sind in Tabelle 6.4 nach einer groben Zuordnung zu linguistischen Informationsebenen gruppiert. Neben inhärenten Eigenschaften der einzelnen Treffer geht auch ihre Relation zu Kontext und Dokumentstruktur in

[7] *Term frequency-inverse document frequency*. Dieses Maß berücksichtigt sowohl die Häufigkeit eines Terminus innerhalb eines Dokuments als auch seine Verteilung innerhalb der gesamten Dokumentkollektion, um eine Aussage über seine Relevanz für das betreffende Dokument zu machen. Vgl. Spärck Jones (1972) und Salton und Buckley (1988).

mehrere der verwendeten Merkmale ein. Dies spiegelt die in Kap. 1 und Kap. 2 ausführlich diskutierte Tatsache wider, dass sich Definitionen in Urteilsbegründungen durch ein hohes Maß an kontextueller und argumentativer Einbindung auszeichnen.

Während die Werte eines Großteils der angeführten Merkmale direkt aus der PreDS der einzelnen Treffer abgelesen werden können, muss in einigen Fällen auf zusätzliche Informationsquellen zugegriffen werden. Neben globalen Dokumenteigenschaften (Anzahl der Sätze, maximale Satzlänge) und manuell zusammengestellten Wortlisten handelt es sich dabei um Synonymie- und Hyperonymieinformation aus Germanet (Hamp und Feldweg (1997)), einem Thesaurus für das Deutsche mit einer vergleichbaren Struktur wie Wordnet (Fellbaum (1998)).

Alle Merkmale lassen sich einfach und effizient berechnen. Aufgrund des heuristischen Charakters der meisten Merkmale, der begrenzten Zuverlässigkeit der PreDS-Analysen und der unzureichenden Abdeckung von Germanet insbesondere im Bereich der juristischen Fachsprache erreicht die Qualität der kodierten Information allerdings nicht in allen Fällen ein optimales Niveau.

### 6.2.3 Ergebnisse verschiedener Klassifikationsverfahren

Auf der Basis der beschriebenen Merkmale haben wir (unter Verwendung der oben beschriebenen Trainingsdaten) eine Auswahl von Klassifizierern verschiedener Typen für die Unterscheidung von *true* und *false positives* unseres dependenzbasierten Definitionsextraktionssystems trainiert. Die trainierten Klassifizierer haben wir dann zur Filterung der Extraktionsergebnisse aller Suchmuster aus dem Goldstandard-Korpus verwendet. Dabei wurden nur die Treffer beibehalten, für die von dem jeweils betrachteten Klassifizierer die Klassenzuweisung [+definitorisch] erzeugt wurde.

#### (a) Verwendete Klassifizierer

Die bei unseren Experimenten verwendeten Klassifizierer sind in folgender Liste kurz beschrieben. Die ersten fünf genannten Klassifizierer repräsentieren die Bandbreite der verschiedenen gängigen Standardansätze, bei den beiden zuletzt angeführten handelt es sich um ein kombiniertes Verfahren und einen sogenannten Metaklassifizierer. Wir können in der Kürze natürlich nur einen groben und unvollständigen Einblick in die wichtigsten Aspekte der Arbeitsweise der jeweiligen Verfahren geben. Für nähere Informationen verweisen wir auf die zu den einzelnen Klassifizierern zitierte Literatur und einschlägige Lehrwerke, z.B. Manning und Schütze (1999):

*Naive Bayes:* Naive Bayes-Klassifizierer berechnen die Wahrscheinlichkeit für jede der zuweisbaren Klassen, gegeben die Werte für die Merkmale einer Instanz, und wählen dann die nach dieser Berechnung wahrscheinlichste Klasse. Der Wahrscheinlichkeitsberechnung liegt ein Modell der bedingten Wahrscheinlichkeiten der jeweiligen Klassen in Abhängigkeit der einzelnen Merkmale zu Grunde. Für die Erzeugung dieses Modells wird die statistische Unabhängigkeit der einzelnen Merkmale untereinander angenommen. Die Werte seiner Parameter können daher einfach und effizient auf der Basis der Verteilung von Klasse und Merkmalen in den Trainingsdaten geschätzt werden. Unsere Experimente stützen sich auf die Naive Bayes-Implementierung aus der WEKA-Toolsuite (Hall u. a. (2009)).

*Entscheidungsbaum:* In Entscheidungsbäumen werden den inneren Knoten eines Baumes Prädikate zugeordnet, die jeweils das Vorliegen eines einzelnen Merkmals der Daten beurteilen. Die Blätter des Baumes repräsentieren Klassenzuordnungen. Bei der Klassifikation passieren Instanzen den Baum vom Wurzelknoten aus und werden an jedem Knoten je nach dem Wert des betreffenden Prädikats entlang einer der ausgehenden Kanten weitergeleitet. Im einfachsten Fall sind alle verwendeten Merkmale binär und jeder Entscheidungsknoten hat entsprechend zwei ausgehende Kanten. Die Klassenzuordnung ergibt sich aus dem jeweils erreichten Blatt des Entscheidungsbaums. Im Trainingsprozess wird ermittelt, wie der Baum strukturiert sein muss und welche Prädikate in welcher Anordnung den einzelnen Knoten zugeordnet werden müssen, um eine möglichst akkurate Klassifizierung der Trainingsinstanzen zu erreichen. Wir verwenden hier den in der WEKA-Toolsuite enthaltenen ADTree-Klassifizierer auf der Basis sogenannter alternierender Entscheidungsbäume (Freund (1999)). In diesen alternieren mit den Entscheidungsknoten sogenannte Vorhersageknoten mit numerischen Werten und einer beliebigen Zahl ausgehender Kanten. Instanzen können bei der Klassifikation unter Umständen alternativ mehrere Pfade passieren. Die Klassenzuordnung wird dann auf der Grundlage der aufsummierten Werte der passierten Vorhersageknoten bestimmt.

*Supportvektor-Maschine:* Supportvektor-Maschinen (SVMs) repräsentieren Trainings- und zu klassifizierende Daten als Punkte in einem multidimensionalen Raum, dessen Dimensionalität sich aus der Anzahl der bei der Kodierung der Daten verwendeten Merkmale ergibt. In der Trainingsphase wird eine Ebene in diesem Raum berechnet, die positive und negative Trainingsdaten optimal voneinander trennt. Bei der Klassifikation neuer Instanzen wird ermittelt, ob der Punkt, der der jeweiligen In-

stanz entspricht, auf der positiven oder negativen Seite dieser Ebene liegt. Wir nutzen die WEKA-Implementierung auf der Basis des SMO-Trainingsverfahrens (*sequential minimal optimization*, Platt (1998)).

*k-Nearest Neighbour:* k-Nearest Neighbour (k-NN)-Klassifizierer repräsentieren ebenfalls Instanzen als Punkte in einem multidimensionalen Raum. Die Klassifizierung neuer Instanzen ergibt sich aus den Klassen einer festgelegten Anzahl in diesem Raum nächstgelegener Trainingsinstanzen (also der k nächsten Nachbarn des neuen Datenpunktes). Wir verwenden hier die an der Universität Tilburg entwickelte k-NN-Implementierung Timbl, in der die Suche nach den bei der Klassifikation relevanten Nachbar-Instanzen durch eine effiziente baumförmige Indexstruktur optimiert wird (Daelemans u. a. (2009), Daelemans und van den Bosch (2005)). In den von uns übernommenen Defaulteinstellung wird für die Klassenzuweisung jeweils nur der nächste Nachbar betrachtet (d.h., $k = 1$), die Distanz zwischen Instanzen ergibt sich aus der Überlappung der Werte ihrer – nach einem Informativitätskriterium gewichteten – Merkmale.

*NBTree:* Der NBTree-Algorithmus (Kohavi (1996)) kombiniert Entscheidungsbäume mit Naive-Bayes Klassifizierern. An den Blättern eines unvollständigen Entscheidungsbaums werden Instanzen mittels Naive-Bayes Klassifizierern bewertet. Wir verwenden hier wieder die in der WEKA-Toolsuite enthaltene Implementierung.

*Boosting:* Ein Boosting-Algorithmus trainiert in mehreren Iterationen schwache sogenannte Basis-Klassifizierer – üblicherweise elementare Entscheidungsbäume – an die nur der Anspruch gestellt wird, dass sie bessere Klassifikationsergebnisse als eine Zufallsauswahl liefern. Nach jeder Iteration werden die Trainingsdaten so umgewichtet, dass fehlklassifizierten Instanzen in der nächsten Iteration ein erhöhtes Gewicht erhalten, sie also stärkeren Einfluss auf das Training des nächsten Basis-Klassifizierers ausüben. Der eigentliche Klassifizierer wird durch lineare Kombination der in den einzelnen Iterationen trainierten Basis-Klassifizierer erzeugt. Wir nutzen den LogitBoost-Klassifizierer (vgl. Friedman u. a. (1998)) aus der WEKA-Toolsuite.

Unter Verwendung des k-NN-Klassifizieres haben wir zusätzlich zwei weitere Experimente durchgeführt, in denen die Trainingsdaten und zu klassifizierende Treffer nicht durch die oben beschriebenen Merkmale, sondern durch lexikalische Vektoren auf der Basis der enthaltenen Lemmata bzw. morphologischen Wurzeln repräsentiert wurden (wie sie auch bei der Ermittlung der

Werte der ähnlichkeitsbasierten Merkmale in Tabelle 6.4 zur Repräsentation von Definiendum und Definiens verwendet wurden). Die Klassifikation in diesen Experimenten erfolgte also rein auf der Basis der lexikalischen Ähnlichkeit zwischen bereits klassifizierten und neuen Treffern.

**(b) Ergebnisse**

In Tabelle 6.5 sind für die verschiedenen Klassifizierer die Präzisions- und Recall-Werte sowie die f-Scores der gefilterten Treffermengen aus dem Goldstandard-Korpus angegeben. Zusätzlich sind als Baseline noch einmal die Werte für die ungefilterten Extraktionsergebnisse angeführt (vgl. Tabelle 5.2).

Um zu erfassen, wie stark die Leistung der verschiedenen Klassifikatoren bei der von uns betrachteten Aufgabenstellung durch die Klassenverteilung in den Trainingsdaten beeinflusst wird (und einen solchen Einfluss gegebenenfalls vermeiden zu können), wurde jedes der Experimente unter zwei verschiedenen Trainingsbedingungen durchgeführt. Diese sind in sind in Tabelle 6.5 unter den Bezeichnungen *unbalanced* und *balanced* gegenübergestellt. Unter der Bedingung *unbalanced* standen dem Klassifizierer in der Lernphase sämtliche 4768 annotierten Trainingsinstanzen zur Verfügung. Die Verteilung des Merkmals [±definitorisch] in dieser Trainingsmenge ist mit 3552 negativen gegen 1216 positive Instanzen stark zu Gunsten des negativen Falls verschoben. Unter der Bedingung *balanced* wurden aus den Trainingsdaten so lange zufällig ausgewählte Negativinstanzen entfernt, bis ein ausgeglichenes Verhältnis zwischen positiven und negativen Instanzen erreicht war.

Das hinsichtlich des f-Score beste Ergebnis wird (mit f=0,33) durch die Filterung mittels eines auf ausgewogenen Daten trainierten NBTree-Klassifizierers erzeugt. Leichte Verbesserungen des f-Score gegenüber einer präzisionsbasierten Suchmusterauswahl ergaben sich in unserem Experiment auch durch Naive Bayes- und SVM-gestützte Klassifikation.

Die Verwendung der *unbalanced*-Trainingsbedingung führt, trotz der insgesamt höheren Anzahl an Trainingsinstanzen, stets zu schlechteren f-Scores als die Verwendung der ausgeglichenen Trainingsmenge. Wie aus Tabelle 6.5 deutlich zu ersehen ist, resultiert die stark erhöhte Proportion negativer Fälle in den Trainingsdaten in einem "vorsichtigeren" Klassifikationsverhalten, also in einer zu Ungunsten des Recall übermäßig erhöhten Präzision bei der Bewertung unbekannter Instanzen. Das beste Ergebnis erzielt hier das Naive Bayes-Verfahren. Der erreichte f-Score von 0,28 entspricht dem für die ungefilterte Ergebnismenge, jedoch kann die Präzision mit 0,5 auf das Zweieinhalbfache erhöht werden.

Die höchste Präzision wird durch SVM-Klassifikation mit unbalancierten Trainingsdaten erreicht. Sie liegt mit 0,59 etwa beim Dreifachen der ungefil-

| | Unbalanced | | | Balanced | | |
|---|---|---|---|---|---|---|
| | *p* | *r* | *f* | *p* | *r* | *f* |
| **Baseline** | | | | | | |
| *(keine Filterung)* | 0,19 | 0,45 | 0,27 | 0,19 | 0,45 | 0,27 |
| **Standardverfahren** | | | | | | |
| *NaiveBayes* | 0,50 | 0,19 | 0,28 | 0,42 | 0,26 | 0,32 |
| *ADTree* | 0,52 | 0,12 | 0,20 | 0,39 | 0,24 | 0,30 |
| *SMO* | 0,59 | 0,13 | 0,21 | 0,42 | 0,25 | 0,31 |
| *k-NN* | 0,36 | 0,16 | 0,23 | 0,28 | 0,26 | 0,27 |
| **Kombinierte / Metaklassifizierer** | | | | | | |
| *NBTree* | 0,56 | 0,16 | 0,25 | 0,38 | 0,29 | 0,33 |
| *LogItBoost* | 0,57 | 0,12 | 0,20 | 0,42 | 0,24 | 0,30 |
| **Lexikalische Klassifikation** | | | | | | |
| *k-NN + Lemmata* | 0,40 | 0,06 | 0,11 | 0,34 | 0,13 | 0,19 |
| *k-NN + Morph. Wurzeln* | 0,30 | 0,11 | 0,17 | 0,27 | 0,13 | 0,18 |

Tabelle 6.5: Klassifikationsergebnisse

terten Extraktion. Allerdings geht der Recall in diesem Fall mit 0,13 auf einen Wert zurück, der zum Beispiel im Kontext der Erschließung einer Rechtsdatenbank kaum noch akzeptabel sein dürfte.

Die schlechtesten Resultate liefert in allen getesteten Szenarios die k-NN Klassifikation.[8] Die noch erheblich geringeren Scores für die lexikalische Klassifikation deuten zudem darauf hin, dass Merkmale auf verschiedenen linguistischen Ebenen (wie diejenigen in Tabelle 6.4) bezüglich des definitorischen

[8] Interessanterweise ergeben sich im Fall der k-NN Klassifikation nur verhältnismäßig geringe Unterschiede zwischen der Performanz unter den beiden unterschiedlichen Trainingsbedingungen. Dies ist wohl dadurch zu erklären, dass Timbl in den von uns übernommenen Standardeinstellungen die Klasse einer Instanz nur anhand eines einzigen nächsten Nachbarn ermittelt. Veränderungen in der Gesamtheit der Trainingsdaten bleiben somit in vielen Fällen ohne Auswirkung auf das Ergebnis.

Status eines Satzes insgesamt deutlich mehr relevante Information kodieren können, als sie allein seiner Lexik entnommen werden kann.

### 6.2.4 Ranking

Wie die Auswertung unserer Experimente zur Filterung mittels verschiedener Klassifizierer gezeigt hat, lässt sich auf diesem Wege eine teils erhebliche Präzisionsverbesserung bei deutlich geringerem Recall-Verlust erzielen als durch die zunächst betrachteten Musterauswahl-Methoden. Jedoch bleibt das oben bereits angesprochene Kernproblem der Ergebnisfilterung ungelöst: Die *true positives*, die zusammen mit den *false positives* aus der Ergebnismenge entfernt werden, stehen in einer Anwendung nicht mehr zur Verfügung.

Dieses Problem lässt sich durch die Erzeugung eines Rankings anstelle einer strikten Filterung beheben. Ein solches Ranking muss hierfür nicht den Anspruch erheben, dass korrekte Treffer auf den oberen Plätzen von höherer Qualität oder Relevanz sind als korrekte Treffer auf den unteren Plätzen (in unserem Fall wäre hierzu zunächst ein vergleichendes Kriterium für die Qualität oder Relevanz von Definitionen genauer auszuarbeiten, als wir dies in Kap. 2 leisten konnten). Ziel ist es hier vielmehr, möglichst viele *false positives* aus der Gesamttreffermenge auf die unteren Plätze zu verschieben. Im Idealfall gelingt es so, die Dichte korrekter Treffer auf den oberen Plätzen so stark zu erhöhen, dass sich in den entsprechenden *top-n*-Segmenten des Ranking eine vergleichbare Ergebnisqualität ergibt wie durch klassifikationsbasierte Filterung. Zugleich bleiben jedoch die weiteren – korrekten wie inkorrekten – Treffer auf den unteren Plätzen des Rankings verfügbar. Sie können somit weiterhin abgefragt werden, wenn dafür sukzessive größere, weniger präzise Ergebnismengen in Kauf genommen werden.

Wir stellen nun eine Ranking-Methode vor, die sich im wesentlichen auf die gleichen Komponenten und Abläufe stützt wie wir sie im vorigen Abschnitt für Klassifikationsverfahren erläutert haben. In Abb. 6.2 ist die entsprechend ergänzte und modifizierte Architektur aus Abb. 6.1 dargestellt.

Ebenso wie für die diskutierten Klassifikationsverfahren wird auch im Falle unseres Ranking-Verfahrens eine Komponente zur Bewertung einzelner Treffer anhand annotierter Daten trainiert. Diese sowie die zu rankenden Treffer werden wiederum durch Merkmale repräsentiert. Wir verwenden hier zunächst ebenfalls die in Tabelle 6.4 angegebenen Merkmale und nehmen dann die Ergebnisse der Klassifikation mit den oben beschriebenen Verfahren als weitere Informationsquelle hinzu.

Anstelle eines Klassifizierers wird in der Lernphase jedoch ein Regressionsmodell mit den Zielwerten 0 bzw. 1 (für definitorische bzw. nicht-definitorische Instanzen) trainiert. Dieses wird direkt zur Errechnung der Scores für die einzel-

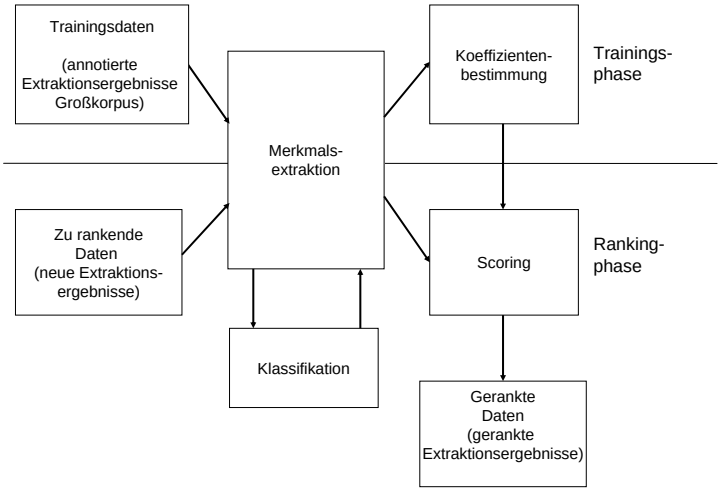



Abbildung 6.2: Komponenten und Informationsfluss beim Ranking

nen Elemente neuer Treffermengen verwendet, nach denen diese dann in einem Ranking sortiert werden. Die Scores stellen somit eine Kombination der durch die Regressionskoeffizienten gewichteten einzelnen Merkmale dar. Die Koeffizienten können als Maß für den Beitrag der einzelnen Merkmale zur Konfidenz in die Korrektheit eines Treffers verstanden werden.

#### (a) Baseline: Ranking nach Präzisionsschätzungen

Als Baseline dient im Folgenden – anknüpfend an die Ergebnisse aus 6.1.1, wo sich Präzisionsschätzungen als das einzige Kriterium zur Musterselektion erwiesen haben, mit dem eine Steigerung des f-Score erzielt werden konnte – ein Ranking, das für jeden Treffer nur die geschätzte Präzision des entsprechenden Suchmusters als Score verwendet. Die Treffer sind somit direkt nach diesen Präzisionsschätzungen sortiert, eine Koeffizienten-Ermittlung durch Regressionsanalyse wird nicht benötigt. Als Grundlage für die Präzisionsschätzungen nut-

zen wir wiederum die annotierte Treffermenge aus dem CORTE-Großkorpus. Relativ zu dieser Baseline evaluieren wir dann ein Ranking auf der Basis des vollen Merkmalssatzes sowie Erweiterungen um verschiedene Kombinationen von Klassifikationsergebnissen. Als Trainingsdaten dienen auch hier die annotierten Treffer aus dem CORTE-Großkorpus.

In Abb. 6.3 sind Präzision und f-Score über dem Recall der (von links nach rechts wachsenden) präzisionsbasiert sortierten Trefferliste abgetragen.

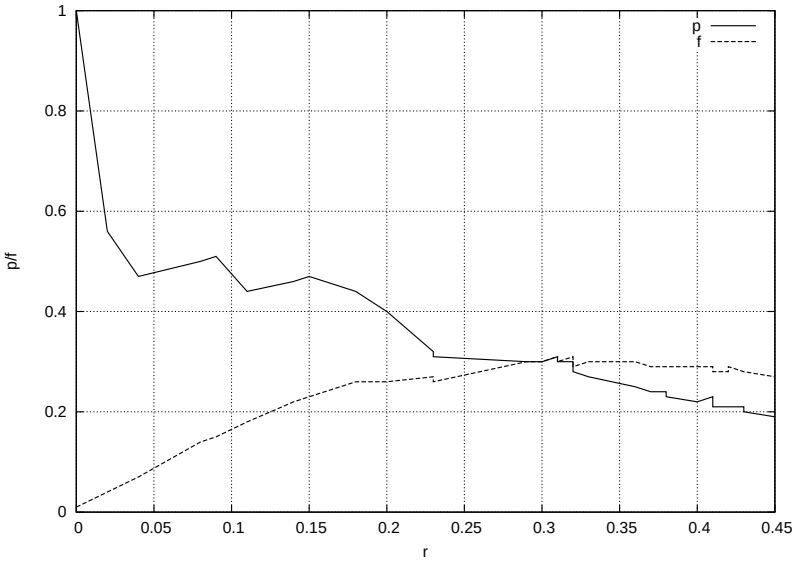



Abbildung 6.3: Präzisions / f-Score-Recall-Graph für das Ranking auf der Basis von Präzisionsschätzungen

Nach dem ersten, korrekten Treffer (aufgrund dessen der Startpunkt des Präzisionsgraphen bei p=1 liegt) fallen die unter dem Ranking erreichten Werte unmittelbar deutlich ab. Der höchste f-Score wird mit 0,31 für p=r=0,31 erreicht.[9]

[9] Die Werte unterscheiden sich von denjenigen, die unsere in 6.1.1 nach denselben Präzisionsschätzungen optimierte Suchmustermenge erzielt (f=0,3, p=0,28, r=0,32). Die Differenz ist darauf zurückzuführen, dass wir dort das Pilotstudien-Korpus für die Musterauswahl nutzen, während wir hier den optimalen f-Score dem Ranking auf dem Goldstandard-Korpus entnehmen. Dies führt u.a. dazu, dass das Top-Segment unseres Ranking mit f=0,31 nur Treffer von 40 der in 6.1.1 ausgewählten insgesamt 46 Suchmuster umfasst.

**(b) Evaluationsmetriken für Rankings**

Die bisher zum Vergleich verschiedener Extraktionsverfahren und Optimierungsmethoden verwendeten Größen Präzision, Recall und f-Score bewerten ungeordnete Ergebnismengen. Um Rankings, also geordnete Ergebnislisten, bewerten und vergleichen zu können, können Präzisions-Recall-Graphen wie in Abb. 6.3 erzeugt werden. Als zusammenfassende Maßzahl lässt sich zunächst einmal (wie oben) der maximale f-Score ermitteln. Daneben haben sich jedoch (insbesondere im Rahmen von *Information Retrieval*-Wettbewerben wie TREC) verschiedene weitere, umfassendere Maße eingebürgert, die aussagekräftigere Vergleiche zwischen mehreren Rankings ermöglichen.[10] Hierzu gehören u.a. Reihen interpolierter Präzisionswerte für bestimmte feste Recall-Levels sowie die *mean average precision* (MAP) der Zielitems. Der Präzisionswert für einen gegebenen Recall-Level wird interpoliert, indem der maximal erreichte Präzisionswert für einen Recall-Score unter dem jeweiligen Ranking ermittelt wird, der den gegebenen Level erreicht oder überschreitet.[11]

Die MAP der Zielitems wird üblicherweise nach folgender Formel[12] ermittelt:

$$MAP = \frac{\sum_{r=1}^{N} p(r) * tp(r)}{|\text{Zielitems in der Grundgesamtheit}|}$$

*N* bezeichnet dabei die Anzahl der gerankten Treffer, p(r) die einem Treffer unter dem Ranking entsprechende nicht-interpolierte Präzision (d.h. die Präzision des *top-n*-Segments bis zu diesem Treffer), und $tp(r)$ eine Funktion mit dem Wert 1 für jedes *true positive r* und 0 für jedes *false positive*.

Im Falle einer Gesamttreffermenge mit einem Recall von 1 ergibt sich die MAP also als Mittelwert der (nicht-interpolierten) Präzisionswerte des Rankings bei jedem korrekten Treffer. Enthält die gerankte Treffermenge nicht alle korrekten Items aus der Grundgesamtheit (r<1, wie in unserem Fall, in dem bereits die zu rankenden Ergebnisse der Definitionsextraktion unvollständig sind),

[10] Zumindest wenn die Rankings nicht den Anspruch einer vergleichenden Relevanzgewichtung der *true positives* erheben, vgl. die obigen Erläuterungen.

[11] Durch diese Definitionen werden die (auch in Abb. 6.3 festzustellenden) "Sägezahn-Muster" von Präzisions-Recall-Graphen ausgeglichen. Diese ergeben sich dadurch, dass der Präzisionswert mit jedem irrelevanten Treffer sinkt, ohne dass der Recall ansteigt, um dann für mehrere aufeinander folgende korrekte Treffer monoton anzusteigen. Folgen in einem Ranking mehrere inkorrekte Treffer dicht aufeinander, fällt der Präzisions-Recall-Graph sprunghaft. Die Definition erlaubt zudem die Interpolation eines Präzisionswertes für den Recall-Level 0, obwohl die Präzision einer Treffermenge mit Recall 0 nicht definiert ist.

[12] Die Definition entspricht der beispielsweise bei der Auswertung der Ergebnisse des TREC 2006 *robust track* verwendeten *average precision over all relevant documents, non-interpolated* (NIST (2006)). Vielfach werden zur Auswertung von Rankings MAP-Varianten ohne Einschränkung auf Zielitems verwendet, die also einen Mittelwert über die Präzision *aller* gerankten Instanzen bilden (z.B. in Pado (2007)).

geht jedes fehlende Item mit p=0 in die Mittelwertberechnung ein (bzw. wird in der angegebenen Formel nur im Nenner berücksichtigt). Der MAP-Wert ist somit kein reines Präzisionsmaß, sondern beinhaltet auch einen Recall-Aspekt.

In Tabelle 6.6 sind für das (in Abb. 6.3 graphisch dargestellte) präzisionsbasierte Ranking der optimale f-Score mit dem zugehörigen Präzisions- und Recall-Wert, die interpolierten Präzisionswerte für Recall-Level zwischen 0 und 0,4 (in Schritten von 0,1) und die MAP angegeben.

| | **Max. f-Score (p, r)** | **Präzision für Recall-Levels** *0* | *0,1* | *0,2* | *0,3* | *0,4* | **MAP** |
|---|---|---|---|---|---|---|---|
| **Ranking** | | | | | | | |
| *präzisionsbasiert* | 0,31 (0,31, 0,31) | 1 | 0,47 | 0,4 | 0,31 | 0,23 | 0,18 |

Tabelle 6.6: Kennzahlen zur Performanz des präzisionsbasierten Rankings

#### (c) Merkmalsbasierte Rankings

Wir vergleichen nun die Ergebnisse von Rankings auf der Grundlage der Merkmale aus Tabelle 6.4 mit den in Tabelle 6.5 zusammengefassten Ergebnissen der klassifikationsbasierten Trefferfilterung. Im einzelnen haben wir folgende Merkmalskonstellationen zur Erzeugung von Rankings verwendet:

1. Die Merkmale aus Tabelle 6.4 bis auf die Präzisionsschätzungen.
2. Alle Merkmale aus Tabelle 6.4.
3. Alle Merkmale und die Ergebnisse des bestbewerteten Klassifikationsvefahrens aus 6.2.3 (Naive Bayes).
4. Alle Merkmale und die Ergebnisse aller Klassifikationsverfahren aus 6.2.3
5. Alle Merkmale und die Ergebnisse der lexikalischen k-NN-Klassifikation

Konstellation (1) nutzt also keine suchmusterspezifische Information. Sie kann für neue Suchmuster ohne aufwändige Annotation weiterer Treffer ermittelt werden und ist daher für eine Übertragung des Ranking-Verfahrens auf andere Suchmustermengen geeignet. Konstellation (2) lässt sich direkt mit den Ergebnissen der klassifikationsbasierten Filterung in 6.2.3 vergleichen. Die Konstellationen (3) bis (5) testen, in wie weit die Klassifikationsergebnisse durch

den Ranking-Mechanismus – trotz der zu erwartenden Abhängigkeit zwischen den Ergebnissen – als Zusatzinformation verwertet werden können.

Für alle fünf Konstellationen haben wir wie in 6.2.3 Trainingsläufe auf der Basis balancierter sowie unbalancierter Daten durchgeführt und die Verwendung linearer und logistischer Regression zur Koeffizienten-Ermittlung erprobt. Die besten Ergebnisse wurden durchgängig durch lineare Regressionsanalyse auf der Grundlage ausgewogener Trainingsdaten erzielt. Tabelle 6.7 fasst die eben erläuterten Maßzahlen für alle Merkmalskonstellationen unter dieser Trainingsbedingung zusammen.

| | **Max. f-Score (p, r)** | **Präzision für Recall-Levels** *0* | *0,1* | *0,2* | *0,3* | *0,4* | **MAP** |
|---|---|---|---|---|---|---|---|
| **Ranking** | | | | | | | |
| *Baseline: präzisionsbasiert* | 0,31 (0,31, 0,31) | 1 | 0,47 | 0,4 | 0,31 | 0,23 | 0,18 |
| *(1) Ling. Merkmale* | 0,35 (0,39, 0,31) | 1 | 0,59 | 0,48 | 0,41 | 0,27 | 0,23 |
| *(2) 1. + P.-Schätzungen* | 0,34 (0,38, 0,31) | 1 | 0,74 | 0,49 | 0,38 | 0,25 | 0,24 |
| *(3) 2. + Naive Bayes* | 0,34 (0,34, 0,33) | 1 | 0,66 | 0,52 | 0,35 | 0,25 | 0,24 |
| *(4) 2. + alle Klassifizierer* | 0,36 (0,39, 0,32) | 1 | 0,65 | 0,5 | 0,42 | 0,28 | 0,25 |
| *(5) 2. + lexikalische k-NN* | 0,37 (0,47, 0,30) | 1 | 0,64 | 0,53 | 0,45 | 0,27 | 0,25 |

Tabelle 6.7: Kennzahlen zur Performanz der merkmalsbasierten Rankings

Sämtliche Rankings führen nach allen betrachteten Maßzahlen zu deutlich besseren Ergebnissen als die rein präzisionsbasierte Baseline. Für die Differenzen der MAP-Werte zur Baseline ergab ein Signifikanztest unter Verwendung des in Yeh (2000) beschriebenen annahmefreien Randomisierungsansatzes in allen Fällen deutlich bessere Signifikanzniveaus als 0,05. Interessanterweise erbringt auch das ohne Präzisionsschätzungen erzeugte Ranking (1) eine klare Verbesserung. Dies deutet darauf hin, dass die von uns zusammengestellten linguistischen Merkmale insgesamt mindestens ebenso informativ bezüglich des definitorischen Werts eines Satzes sind wie die auf der Basis von Expertenwissen ermittelten Suchmusterpräzisionen.

Die Unterschiede zwischen den merkmalsbasierten Rankings fallen hingegen relativ gering aus und lassen kein eindeutiges Urteil über deren relative

Qualität zu.[13] So führt das aus linguistischen Merkmalen und Präzisionsschätzungen erzeugte Ranking (2) nach dem Großteil der betrachteten Maße nicht zu den besten Ergebnissen, scheint aber (zumindest im von uns betrachteten Fall) die beste Treffersortierung im obersten Segment zu leisten. Dagegen liefert das insgesamt bestbewertete Ranking (5) in diesem Bereich eine relativ geringe Optimierung.

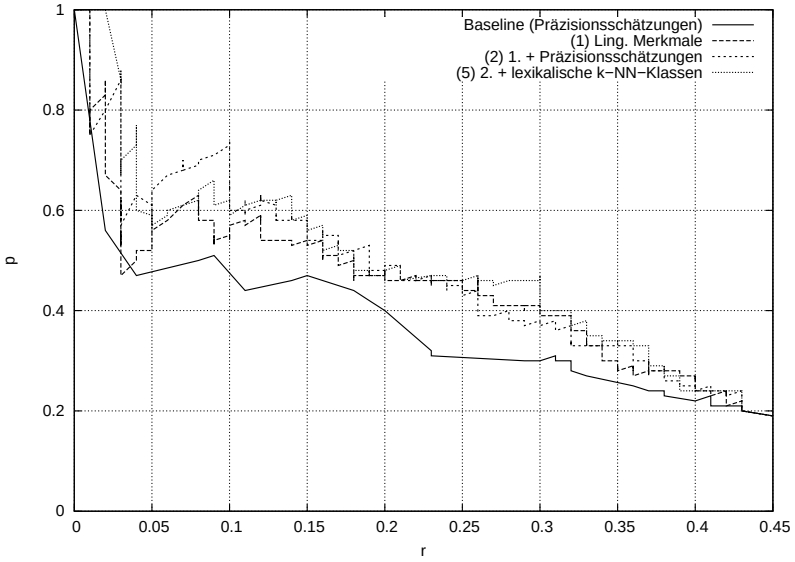



Abbildung 6.4: Präzisions-Recall-Graphen für merkmalsbasierte Rankings

In Abb. 6.4 sind für die Rankings (1), (2) und (5) Präzisions-Recall-Graphen gezeigt und der im vorigen Abschnitt diskutierten Baseline gegenübergestellt.[14] Das Diagramm deutet ebenso wie die Daten in Tabelle 6.7 darauf hin, dass sich Ranking (1) durch eine gute Performanz bei niedrigem Recall auszeichnet, während Ranking (5) eine höhere Stabilität im mittleren Bereich aufweist. Insgesamt lassen sich jedoch auch der graphischen Darstellung keine klareren

[13] Die Differenzen zwischen den merkmalsbasierten Rankings stellten sich als (nach gängigen Maßstäben, d.h. $\alpha \geq 0.05$) nicht signifikant heraus. Das beste Signifikanzniveau ergab sich hier mit 0.1 für die Differenz zwischen den Rankings (1) und (5).

[14] Aus Gründen der Übersichtlichkeit haben wir im Gegensatz zu Abb. 6.3 auf die entsprechenden f-Score-Diagramme verzichtet

Ergebnisse bezüglich des Unterschieds zwischen den merkmalsbasierten Rankings entnehmen als der Zusammenfassung in Tabelle 6.7.

#### (d) Analyse der ermittelten Merkmalsgewichtung

Die durch die Regressionsanalyse ermittelten Merkmalsgewichte dienen uns hier in erster Linie als Koeffizienten der Scoring-Funktion. Sie beinhalten aber gleichzeitig auch Information darüber, in welchem Maße die einzelnen gewichteten Merkmale insgesamt zum Definitionscharakter eines Satzes beitragen. Ein Vergleich der Gewichte erlaubt also eine Validierung der intuitiven Annahmen, von denen wir bei der Auswahl und Formulierung der einzelnen Merkmale ausgegangen sind.

In Tabelle 6.8 sind für die Merkmalskonstellationen (1), (2) und (5) die jeweils höchst- und niedrigstgewichteten Merkmale (in absteigender Reihenfolge) wiedergegeben. Die in der Tabelle nicht angegebenen Merkmale wurden mit Werten gleich oder nahe 0 bewertet. Auffällig ist insofern, dass nur ein relativ geringer Teil der Merkmale aus Tabelle 6.4 überhaupt eine Gewichtung im relevanten Bereich erhalten hat. Für einige der weggefallenen Merkmale (insbesondere die auf Germanet-Information basierenden) dürfte dies daran liegen, dass sie nur für einen geringen Teil der Trainings- und Testinstanzen ermittelt werden konnten oder (wie im Falle der Rechtsprechungszitate) die bei ihrer Ermittlung verwendete Information nicht zuverlässig genug war.

Die relativen Gewichtungen der angegebenen Merkmale entsprechen allerdings in etwa unseren Erwartungen. Die relativ hohe Gewichtung des Ergebnisses der lexikalischen Klassifikation in (5) korrespondiert zu der Beobachtung, dass durch Hinzunahme dieses Merkmals die besten Ranking-Ergebnisse erzielt werden. Die Tatsache, dass die morphologische und lexikalische Ähnlichkeit der Definitionsbestandteile stark unterschiedliche Gewichte erhalten, führen wir darauf zurück, dass nur die Zerlegung bzw. Zusammenfügung von Komposita (wie sie durch die morphologische Ähnlichkeit erfasst wird) einen wirklich definitionsrelevanten Prozess darstellt, während die bloße Verwendung inhaltlich verwandter Begriffe als generelles Mittel der Kohärenzerzeugung in den verschiedensten Aussagetypen dient.

#### (e) Ranking der Extraktionsergebnisse aus dem CORTE-Großkorpus

Um die Leistungsfähigkeit des dargestellten Ansatzes zum Ergebnisranking auch anhand größerer Datenmengen beurteilen zu können, haben wir zusätzlich zu den bisher diskutierten Experimenten auch für die Extraktionsergebnisse aus dem CORTE-Großkorpus die Ranking-Methoden (1) und (2) erprobt. Da eine gleichzeitige Nutzung aller annotierten Treffer als Trainings- und Testdaten die

| *Gewichtung* | *Konstellation (1)* | *Konstellation (2)* | *Konstellation (5)* |
|---|---|---|---|
| *positiv* | | | |
| ↑ | Definiendum vor Definiens | Präzisionsschätzung | Präzisionsschätzung |
| ↑ | Anaphorisches Definiendum | morph. Ähnlichkeit der Bestandteile | morph. Ähnlichkeit der Bestandteile |
| ↑ | definiter Genusbegriff | Definiendum vor Definiens | Definiendum vor Definiens |
| ↑ | Treffer im Kontext | definiter Genusbegriff | Anaphorisches Definiendum |
| ↑ | morphol. Ähnlichkeit der Bestandteile | Anaphorisches Definiendum | lexikalische Klassifikation |
| ↑ | Boostwort | Treffer im Kontext | definiter Genusbegriff |
| ↑ | Modifikator | TF / IDF des Definiendum | Treffer im Kontext |
| ↑ | TF / IDF des Definiendum | | Boostwort |
| **0** | | | |
| ↓ | relative Dokumentposition | relative Dokumentposition | Definition kein Hauptsatz |
| ↓ | Definition kein Hauptsatz | Definition kein Hauptsatz | Definites Definiendum |
| ↓ | Subsumtionssignal | Subsumtionssignal | Subsumtionssignal |
| ↓ | lexikalische Ähnlichkeit der Bestandteile | Definites Definiendum | Negation |
| ↓ | Definites Definiendum | Negation | lexikalische Ähnlichkeit der Bestandteile |
| ↓ | Stoppwort (automatische Erweiterung) | lexikalische Ähnlichkeit der Bestandteile | Stoppwort (automatische Erweiterung) |
| ↓ | Negation | Stoppwort (automatische Erweiterung) | |
| *negativ* | | | |

Tabelle 6.8: Relative Gewichtung der Merkmale (in verschiedenen Konstellationen)

Aussagekraft der Ergebnisse stark beeinträchtigen würde, haben wir das Experiment fünffach kreuzvalidiert durchgeführt: Die annotierten Treffer wurden in fünf verschiedenen Einteilungen (jeweils im Verhältnis 4:1) partitioniert. Die größere Partition wurde jeweils als Trainingsdaten (sowie zur Schätzung der Suchmusterpräzisionen) und die kleinere zur Evaluation verwendet.

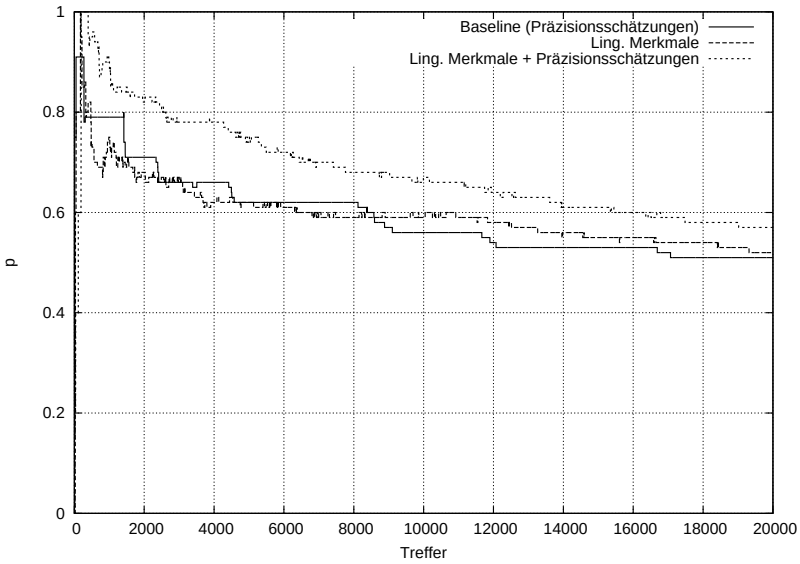



Abbildung 6.5: Ranking der Treffer aus dem CORTE-Großkorpus (top 20 000, gemittelt und extrapoliert)

Weil wir für die Extraktion aus dem CORTE-Großkorpus keine Angaben zum Recall machen können, können die Rankings der Extraktionsergebnisse auch nicht auf der Basis von Präzisions-Recall-Graphen und den bisher verwendeten Maßzahlen verglichen werden. Eine Evaluation kann nur bezüglich absoluter Trefferzahlen vorgenommen werden. In Abb. 6.5 sind die – über die fünf Training-Test-Einteilungen gemittelten – Präzisionen für das präzisions- und mekmalsbasierte Ranking über wachsenden Trefferzahlen aufgetragen (also die Präzision für entsprechend große *n-best*-Listen). Dargestellt ist das obere Ranking-Segment bis zum zehntausendsten Treffer. Die Präzisionswerte wurden von den annotierten auf die unannotierten Treffer extrapoliert, indem jedem nicht annotierten Treffer der Präzisionswert des Ranking beim nächstniedrigeren annotierten Treffer zugewiesen wurde.

Tabelle 6.9 enthält die mittleren interpolierten Präzisionswerte bei 10, 100, 500, 1000, 5000 und 20 000 Treffern sowie den mittleren *bpref*-Score der Rankings. Bei diesem Wert handelt es sich um eine zusammenfassende Maßzahl für Rankings, in denen nur für einen geringen Teil der sortierten Treffer bekannt ist, ob es sich um *true positives* oder *false positives* handelt. Betrachtet wird, wie viele bekannte *false positives* durch das gegebene Ranking vor bekannte *true positives* sortiert werden. Die Berechnung erfolgt nach folgender Formel (vgl. NIST (2006)):

$$\text{bpref} = \frac{1}{R}\sum_{r}(1 - \frac{|\{n : s(n) > s(r)\}|}{min(R,N)})$$

Dabei steht $R$ für die Anzahl bekannter *true positives*, $N$ für die Anzahl bekannter *false positives*, $s$ bezeichnet die verwendete Scoring-Funktion, $r$ iteriert über die bekannten *true positives* und $n$ über die $R$ bestgerankten *false positives*.

| | **Präzision für n beste Treffer** | | | |
|---|---|---|---|---|
| | *10* | *100* | *500* | *1000* |
| **Ranking** | | | | |
| *Baseline* | 0,92 | 0,92 | 0,80 | 0,80 |
| *(1) Ling. Merkmale* | 1 | 1 | 0,84 | 0,81 |
| *(2) Ling. Merkmale + Präzisionsschätzungen* | 1 | 1 | 0,97 | 0,92 |

| | **Präzision für n beste Treffer** | | **bpref** |
|---|---|---|---|
| | *5000* | *20 000* | |
| **Ranking** | | | |
| *Baseline* | 0,62 | 0,51 | 0,48 |
| *(1) Ling. Merkmale* | 0,64 | 0,53 | 0,47 |
| *(2) Ling. Merkmale + Präzisionsschätzungen* | 0,76 | 0,57 | 0,55 |

Tabelle 6.9: Ranking der Extraktionsergebnisse aus dem CORTE-Großkorpus

Sowohl die graphische Darstellung als auch die zusammenfassenden Werte in Tabelle 6.9 lassen eine deutliche Verbesserung durch das Ranking 2 (nach linguistischen Merkmalen und Präzisionsschätzungen) erkennen. Die für die Extraktionsergebnisse aus dem Goldstandard-Korpus festgestellte gute Performanz des rein auf linguistischen Merkmalen basierenden Rankings (Ranking 1

in Tabelle 6.9) ist auf der erweiterten Datengrundlage nicht in vollem Maße reproduzierbar, allerdings ist ein direkter Vergleich aufgrund des Fehlens exakterer Recall-Werte für die Extraktion aus dem CORTE-Großkorpus nicht möglich. Ranking 1 liegt zwar an den betrachteten Fixpunkten über dem präzisionsbasierten Baseline-Ranking, erzielt jedoch einen geringfügig niedrigeren bpref-Wert, was auf eine insgesamt schlechtere Ergebnisqualität in dem in Abb. 6.5 nicht mehr dargestellten unteren Ranking-Segment hindeutet.

Insgesamt erzielt die dependenzbasierte Definitionssuche im CORTE-Korpus 108 209 Treffer und eine Gesamtpräzision von 0,2 (siehe 5.3). Somit werden durch das optimale Ranking 2 deutlich mehr als die Hälfte aller in der Treffermenge vorhandenen Definitionen noch in einen Bereich mit einer Präzision von 0,5 und mehr sortiert. Geht man als grobe Schätzung wiederum (vgl. ebenfalls 5.3) von ca. 59 000 Definitionen im CORTE-Korpus aus, entspricht das knapp einem Fünftel aller im Korpus vorhandenen Definitionen.

Insbesondere für Anwendungen, die Rankinginformation direkt nutzen können, steht mit dem hier dargestellten Verfahren somit eine Optimierungsmöglichkeit zur Verfügung, die die praktische Einsetzbarkeit des CORTE-Systems erheblich steigert. Das Rankingverfahren kann beispielsweise zur Treffersortierung in einer Suchmaschine für den stichwortbasierten Zugriff auf Definitionen genutzt werden. Unsere Evaluation lässt erwarten, dass so Resultatlisten mit einer hohen Dichte brauchbarer Treffer auf den ersten fünfzig bis hundert Plätzen erzeugt werden können (die im Normalfall für menschliche Benutzer besonders relevant sind). Um dies zu bestätigen, muss allerdings noch näher untersucht werden, ob das Rankingverfahren in allen Fällen gleichmäßig effektiv arbeitet. Beispielsweise ist zu ermitteln, in wie weit über Definitionen verschiedener Begriffe und Definitionstypen hinweg stabile Verbesserungseffekte erzielt werden. Wir können eine solche detaillierte Evaluation hier nicht mehr vornehmen. In Kap. 7 gehen wir kurz auf einen Prototypen einer Definitionssuchmaschine auf der Basis des CORTE-Systems und -Korpus ein, mit dessen Hilfe unsere Evaluation um eine begriffszentrierte Auswertung in einem *end-to-end*-Szenario erweitert werden kann.

## 6.3 Bootstrapping

Filterungs- und Ranking-Methoden, wie wir sie in den vorigen Abschnitten diskutiert haben, sind geeignet, um Treffermengen mit erhöhter Präzision zu erzeugen. Dagegen kann der Recall einer Informationssuche (zumindest innerhalb des Rahmens eines musterbasierten Verfahrens) nur durch die Hinzunahme weiterer Extraktionsmuster vergrößert werden. Dem in dieser Arbeit bisher verfolgten Ansatz der manuellen Muster-Spezifikation aufgrund einer intellek-

tuellen Datenanalyse sind hierbei Grenzen gesetzt. Sowohl die Annotation der benötigten Datenmenge als auch die Umsetzung annotierter Daten in Suchmuster sind extrem arbeitsaufwändige Vorgänge.

Da sich dieses Problem im Rahmen fast aller Aufgaben im Bereich des textbasierten Informationszugriffs stellt, galt in den letzten Jahren besonderes Forschungsinteresse induktiven Methoden zum automatischen korpusbasierten Erwerb von Suchmustern. Neben Methoden zum direkten automatischen Erwerb von Extraktionsregeln kommen vermehrt *Bootstrapping*-Techniken zum Einsatz (z.B. Riloff und Jones (1999); Xu (2007)). Mit diesen können iterativ aus einer initialen Menge von Beispieltreffern (sog. Seeds) und einem großen Korpus Suchmustermengen erzeugt und konsolidiert werden. *Bootstrapping* vermeidet somit die Abhängigkeit von der Verfügbarkeit und Qualität manuell annotierter Trainingsdaten.

Wir stellen im Folgenden die Ergebnisse vor, die wir in einem Experiment zur Erzeugung von Mustern für die PreDS-basierte Definitionsextraktion mittels eines Bootstrapping-Verfahrens erzielen konnten. Bei dem dargestellten Verfahren handelt es sich um eine verhältnismäßig einfache Umsetzung des Bootstrapping-Ansatzes, die an einigen Stellen Raum für offensichtliche Verbesserungen bietet. Dennoch konnten wir durch die Kombination der automatisch erworbenen Suchmuster mit den manuell spezifizierten und dem eben erläuterten Ranking-Verfahren für die Extraktion aus dem Goldstandard-Korpus bereits eine signifikante Ergebnisverbesserung erreichen.

### 6.3.1 Verfahren

#### (a) Typischer Bootstrapping-Prozess

Ein Bootstrapping-Prozess, wie er etwa zum Erwerb von Suchmustern für die Nutzung in IE-Systemen eingesetzt wird, verläuft zyklisch. Jeder Zyklus besteht typischerweise aus folgenden Schritten (vgl. Agichtein und Gravano (2000)):

1. *Seed-Auswahl:* Es wird eine Menge sog. *Seeds* gebildet. Je nach Komplexität der IE-Aufgabe handelt es sich hierbei entweder um Instanzen des gesuchten Informationstyps (etwa Eigennamen aus einer bestimmten Klasse) oder ausgewählte Schlüsselwörter aus solchen Instanzen (etwa Beispiele für die Relata einer Relation, z.B. die Namen von Funktionsinhabern und ihren Nachfolgern für die im Rahmen von MUC-6 eingeführte *management succession task*).

2. *Suche anhand der Seeds:* Die Seeds werden in einem großen, thematisch geeigneten Textbestand gesucht und gegebenenfalls nach bestimmten Kriterien Treffer ausgewählt.

3. *Erzeugung der Suchmusterkandidaten:* In diesen Treffern werden, ausgehend von den Auftreten der Seeds, Suchmusterkandidaten identifiziert.

4. *Suchmusterauswahl:* Aus den Suchmusterkandidaten werden (u.U. auch unter Berücksichtigung der bereits vorhandenen Suchmuster) diejenigen ausgewählt, die beibehalten werden sollen.

5. *Suche anhand der neuen Muster:* Mit den erzeugten Suchmustern werden Instanzen des gesuchten Informationstyps extrahiert. Diese werden (vollständig oder nach Durchlaufen einer Auswahlprozedur) wiederum als Quelle für die Seeds des nächsten Zyklus verwendet.

Der Zyklus wird wiederholt, bis eine vorgegebene Abbruchbedingung erfüllt ist. Der dargestellte Prozess erlaubt in allen Schritten eine große Bandbreite an Realisierungsmöglichkeiten. So kann sowohl die Seed- als auch die Suchmusterauswahl mittels unterschiedlichster Methoden (einschließlich der Einbindung von Experten für die Durchsicht der Suchmusterkandidaten) erfolgen. Die Abbruchbedingung kann etwa als feste Maximalzahl, im Hinblick auf Veränderungen der erzeugten Suchmustermenge oder mit Bezug auf Eigenschaften der Treffer der erzeugten Suchmuster formuliert werden. Die Art und Granularität der verwendeten Seeds sowie die Methode, nach der aus den Fundstellen der Seeds im Korpus in Schritt (3) Suchmusterkandidaten erzeugt werden, sind dagegen durch den zu extrahierenden Informationstyp und die Korpusrepräsentation mehr oder weniger stark determiniert.

#### (b) Umsetzung für unser Experiment

Bootstrapping-Verfahren wurden zunächst mit Erfolg vor allem für Information Extraction-Aufgaben von verhältnismäßig geringer Komplexität (Named Entity- oder Relationserkennung) eingesetzt, bei denen die Textgrundlage ungeparst oder mit Analysen in Form von Dependenztupeln vorlag. In solchen Fällen können typischerweise einzelne Substantive als Seeds genutzt werden. Auf deren Basis erzeugt der Bootstrapping-Prozess dann jeweils direkte Kontexte oder unmittelbar übergeordnete Verben als neue Muster-Vorschläge.

Bei Definitionsbestandteilen handelt es sich allerdings in den seltensten Fällen um einzelne, dem Definitionsprädikat direkt untergeordnete Wörter, und die im CORTE-System genutzten Suchmuster beschreiben hierarchische Konfigurationen innerhalb baumförmiger Strukturen (den PreDS Strukturen, die wir

als Textrepräsentationen verwenden).[15] Für den Einsatz eines Bootstrapping-Verfahrens im Kontext einer solchen Konstellation existiert bisher keine paradigmatische Vorgehensweise zur Seed- und Musterkandidaten-Erzeugung. Das Hauptaugenmerk beim Entwurf unserer Bootstrapping-Komponente galt daher den oben unter (1) und (3) beschriebenen Schritten. Für die restlichen Schritte haben wir möglichst einfache Umsetzungen gewählt. Im einzelnen liegt dem Bootstrappingverfahren in unserem Experiment folgender Ablauf zu Grunde:

1. *Seed-Auswahl:* Als initiale Seeds haben wir manuell ausgewählte Paare von Inhaltswörtern (je ein Wort aus Definiens und Definiendum) aus einer für Lernzwecke zusammengestellten Sammlung von ca. 140 juristischen Definitionen verwendet.[16] Seeds für weitere Iterationen wurden aus den Paaren der Kopf-Lemmata der durch die neuen Suchmuster identifizierten Definiendum- und Definiens-Phrasen gewonnen. Hierfür wurden solche Begriffe ausgewählt, deren gemeinsames Auftreten möglicherweise durch den Inhalt bestimmter Definitionen bedingt ist. Die Begriffspaare wurden hierfür nach relativem Assoziationsgrad ausgewählt. Dieser wurde durch den Vergleich von Log-Likelihood Rankings der Paare auf der Grundlage der Frequenzen in den extrahierten Sätzen und im CORTE-Großkorpus ermittelt. Ausgewählt wurden Paare, die in den extrahierten Definitionen erheblich höher gerankt wurden als im CORTE-Großkorpus.

2. *Suche anhand der Seeds:* Gesucht wurde im CORTE-Großkorpus auf Lemmabasis nach allen Sätzen, die beide Elemente eines Seed-Paares als Geschwisterknoten unterhalb eines (beliebig weit) übergeordneten Prädikats enthalten.

3. *Erzeugung der Suchmusterkandidaten:* Es wurden die allgemeinsten Suchmusterkandidaten identifiziert, die gemeinsam die in Schritt (2) erhaltenen Treffer erzielen können. Für jeden extrahierten Satz wurde zunächst das dem Seed-Paar übergeordnete Prädikat mit den Pfaden zu den beiden Seed-Elementen identifiziert. Als Suchmusterkandidaten wurden dann für jedes Prädikat die kürzesten Pfadpaare ausgewählt, die von keinem anderen Pfadpaar mit demselben Prädikat subsumiert wurden (für die also kein weiteres Pfadpaar mit demselben Prädikat existiert, dessen Pfade die entsprechenden Pfade des ausgewählten Paars als Präfix enthalten).

[15] Vgl. Stevenson und Greenwood (2006) zu einem generischen Modell der Komplexität des Suchmustererwerbs für verschiedene Typen von Repräsentationen linguistischer Strukturen

[16] `http://www.jurawiki.de/JuristischeDefinition/AlleDefinitionen`, 27.11.2007

4. *Suchmusterauswahl:* Als Suchmuster wurden dann die 110 Muster mit dem höchsten Support in der Treffermenge aus Schritt (2) beibehalten, d.h. diejenigen, mit denen sich die größte Zahl der in dieser Menge enthaltenen Treffer erzielen lässt. Diese wurden durchgesehen und ggf. um Attribute für Tempus und Genus Verbi ergänzt.

5. *Suche anhand der neuen Muster:* Die Muster wurden in das von unserem Extraktionssystem verwendete Format gebracht und auf das CORTE-Großkorpus angewendet.

Die Abbruchbedingung haben wir für unser Experiment im Vorhinein ohne Bezug auf die gewonnenen Suchmuster bzw. deren Treffer festgesetzt: Der Zyklus wurde zweimal durchlaufen.

### 6.3.2 Ergebnisse

Wir gehen nun auf die Ergebnisse der Anwendung des eben beschriebenen Bootstrapping-Prozesses ein. Wir betrachten drei Szenarios:

(a) Der Bootstrapping-Zyklus wird nur einmal durchlaufen

(b) Der Zyklus wird zweimal durchlaufen. Die im ersten Durchlauf gewonnenen Suchmuster werden im zweiten Durchlauf zunächst verworfen, neue Suchmuster werden dann aus den Kandidaten beider Durchläufe (auf der Basis der aufsummierten Support-Werte aus beiden Durchläufen) ausgewählt.

(c) Im zweiten Durchlauf werden 110 zusätzliche Muster erzeugt.

Wir untersuchen zunächst die erzeugten Suchmuster und deren Performanz im Vergleich zu den bisher verwendeten manuell erstellten Extraktoren. Anschließend kombinieren wir beide Mustermengen und optimieren die Ergebnisse unter Verwendung des in 6.2.4 beschriebenen Ranking-Verfahrens.

#### (a) Extrahierte Muster

Bereits nach dem ersten Durchlauf des oben dargestellten Zyklus erzeugt unser Bootstrapping-Prozess plausible Suchmuster, einschließlich einiger, die identisch oder ähnlich auch in unserem manuell spezifizierten Musterbestand enthalten sind. Bei einer detaillierten Inspektion der Ergebnisse des ersten Zyklus haben wir in 23 Fällen eine vollständige Subsumtionsbeziehung zwischen einem manuell spezifizierten und einem generierten Muster festgestellt, in einigen weiteren Fällen eine deutliche Ähnlichkeit. Zugleich wurden durch den

Bootstrapping-Prozess jedoch insgesamt 31 Prädikate (zum Teil mit mehreren Valenzrahmen) identifiziert, die in keinem der bereits vorhandenen Suchmuster auftauchen.

Andererseits enthalten die Ergebnisse allerdings auch ein hohes Maß an "Rauschen". Dieses setzt sich zusammen aus:

1. Offenkundig unbrauchbaren Mustern (inklusive solcher, die aus inkorrekten PreDS gewonnen wurden)
2. Mustern, die mit hoher Wahrscheinlichkeit zu sehr unpräzisen Ergebnissen führen (insbesondere verschiedene kaum eingeschränkte Kombinationen mit *sein*)
3. Muster mit unplausiblen *frames* (die also höchstwahrscheinlich keine korrekte Identifikation von Definitionsbestandteilen ermöglichen, jedoch dennoch stark mit korrekten Definitionsmustern korreliert sein können)

In Tabelle 6.10 ist für die Ergebnisse des ersten Bootstrapping-Durchlaufs eine Einteilung in plausible Kandidaten sowie Problemfälle der drei genannten Kategorien angegeben und mit typischen Beispielen illustriert. Für den zweiten Durchlauf (in den oben beschriebenen Szenarios (b) und (c)) haben wir nur die hinzugekommenen Prädikate erfasst und bewertet. Die Ergebnisse sind ebenfalls in Tabelle 6.10 wiedergegeben.

#### (b) Performanz der erzeugten Suchmuster

Für die Bewertung der Extraktionsleistung der neu gewonnenen Suchmuster haben wir die Gesamtperformanz für die drei Suchmustermengen aus einem Zyklus, zwei Zyklen nach Methode (a) und zwei Zyklen nach Methode (b) anhand unseres Goldstandard-Korpus ermittelt. Zusätzlich habe wir Rankings aller drei Extraktionen nach dem in 6.2.4 als optimal ermittelten Setup erzeugt[17] und schließlich die Ergebnisse einer Kombination mit dem manuell spezifizierten Suchmusterbestand untersucht.

In Tabelle 6.11 sind die Scores für die drei allein durch Bootstrapping gewonnenen Suchmustermengen zusammengefasst. Abb. 6.6 zeigt Präzisions-Recall-Graphen für die Rankings der Suchergebnisse.

Der Vergleich der drei untersuchten Bedingungen lässt erkennen, dass – zumindest in der von uns erprobten Umsetzung – die Ergebnisqualität des ersten

[17] Die für das Ranking benötigten Präzisionsschätzungen für die einzelnen Muster konnten nur anhand der bereits annotierten Treffer aus dem CORTE-Großkorpus erzeugt werden. Nur dann, wenn für ein Suchmuster weniger als 20 Treffer annotiert waren, wurden weitere Treffer bis zum Erreichen dieser Grenze inspiziert. Es ist davon auszugehen, dass durch diese Vorgehensweise nur ungenaue und verzerrte Schätzungen erzeugt werden konnten.

**1. Durchlauf (Auswertung der Suchmuster)**

| | *Anzahl* | *Beispiele* |
|---|---|---|
| *Plausibel* | 46 | *bestehen* + Subjekt + Konditionalsatz<br>*bestehen* + Subjekt + *in*-PP<br>*genügen* + Subjekt + *für*-PP<br>*erkennen* + direktes Objekt + *in*-PP |
| *Unbrauchbar* | 23 | *sein* + indirektes Objekt<br>*ausschließen* + direktes Objekt + *mit*-PP<br>*bebauen* + direktes Objekt + *in*-PP |
| *Unpräzise* | 19 | *sein* + Subjekt + Prädikatsnomen |
| *Unplausibler frame* | 22 | *voraussetzen* + Subjekt + *gegen*-PP |

**2. Durchlauf (Auswertung der neuen Prädikate)**

| | *Methode (a)* | *Methode (b)* | *Beispiele* |
|---|---|---|---|
| *Veränderung zum ersten Durchlauf* | +10 / -16 | +20 | |
| *Plausibel* | 7 | 12 | *bedeutsam, betreffen, beurteilen, entsprechen* |
| *Unplausibel* | 3 | 8 | *treffen, auskönnen, begegnen, erwachsen* |

Tabelle 6.10: Inspektion der durch Bootstrapping gewonnenen Suchmuster

Bootstrapping-Zyklus durch Einbeziehung weiterer Durchläufe nicht mehr gesteigert wird. Der in Abb. 6.6 erkennbare Zugewinn einiger weniger Ergebnisse bei voller Präzision schlägt sich im Gegensatz zum sonstigen Präzisionsverlust in keiner der zusammenfassenden Maßzahlen nieder. Es zeigt sich jedoch auch, dass mit den im ersten Zyklus erworbenen Suchmustern eine Ergebnisqualität erzielt wird, die im oberen Segment des Rankings im selben Bereich liegt wie diejenige der manuell spezifizierten Muster.

Die Ergebnisse der Extraktion mit der Kombination beider Suchmusterbestände deuten darauf hin, dass dies nicht allein auf die Überschneidung zwi-

**Gesamtergebnisse**

| | *p* | *r* | *f* |
|---|---|---|---|
| **Zyklen** | | | |
| *1* | 0,15 | 0,5 | 0,23 |
| *2 (Methode b)* | 0,13 | 0,53 | 0,22 |
| *2 (Methode c)* | 0,12 | 0,58 | 0,2 |

**Ranking**

| | **Max. f-Score (p, r)** | **Recall Level** *0* | *0,1* | *0,2* | *0,3* | *0,4* | **MAP** |
|---|---|---|---|---|---|---|---|
| **Zyklen** | | | | | | | |
| *1* | 0,32 (0,33, 0,30) | 1 | 0,67 | 0,38 | 0,33 | 0,22 | 0,21 |
| *2 (Methode b)* | 0,30 (0,37, 0,25) | 1 | 0,44 | 0,40 | 0,31 | 0,22 | 0,20 |
| *2 (Methode c)* | 0,25 (0,18, 0,38) | 1 | 0,44 | 0,24 | 0,18 | 0,18 | 0,17 |

Tabelle 6.11: Ergebnisse der Definitionssuche mit den durch Bootstrapping erzeugten Suchmustern

schen dem manuell spezifizierten und dem durch das Bootstrapping-Verfahren erworbenen Suchmustersatz bedingt ist. Tabelle 6.12 gibt wiederum Präzision, Recall und f-Score für die nicht gerankten Suchergebnisse aus dem Goldstandard-Korpus an und enthält außerdem die zusammenfassenden Maße für das Ranking nach Präzisionsschätzungen und lexikalischen Merkmalen. Dieses ist zusätzlich in Abb. 6.7 als Graph dem optimalen Ranking der ohne Bootstrapping erzielten Suchergebnisse und der Baseline aus 6.2.4 (präzisionsbasiertes Ranking derselben Ergebnismenge) gegenübergestellt. Für eine parallele Auswertung der Rankings auf dem CORTE-Großkorpus (wie wir sie für die gerankten Ergebnisse der manuell spezifizierten Suchmuster in 6.2.4 diskutiert haben) wäre zunächst die Annotation größerer Stichproben der Extraktionser-

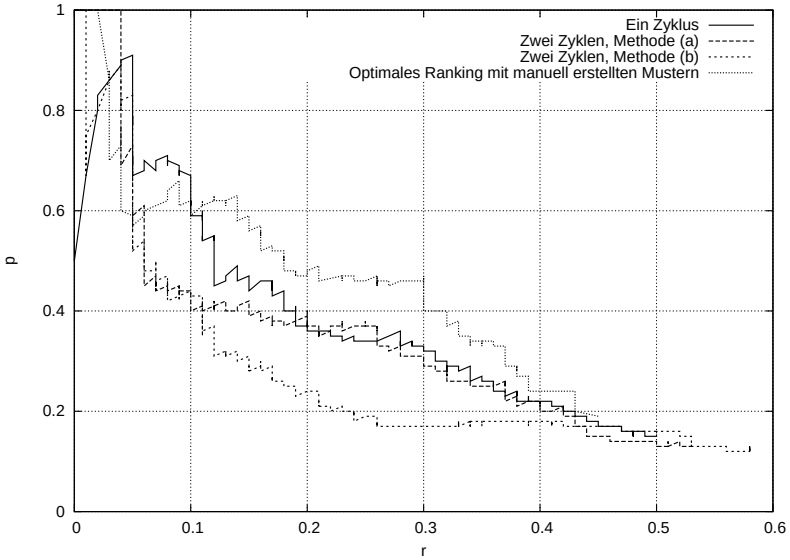



Abbildung 6.6: Ranking der Extraktionsergebnisse der durch Bootstrapping erzeugten Suchmuster

gebnisse der automatisch erzeugten Suchmuster notwendig gewesen.[18] Diese konnte im Rahmen der hier vorgestellten Arbeiten nicht geleistet werden.

Die Ergänzung unseres Suchmusterbestandes um die durch Bootstrapping erzeugten Extraktoren führt zu einem deutlich (um 18 Prozentpunkte) höheren Recall-Zuwachs als die manuelle Spezifikation weiterer Suchmuster in 6.1.2. Für die nicht gerankten Suchergebnisse liegt der f-Score jedoch aufgrund des Präzisionsverlusts von weiteren 3 Prozentpunkten unter dem dort erreichten. Dieser Effekt kann durch die Anwendung des im vorigen Abschnitt entwickelten Ranking-Verfahrens aufgefangen werden.

Das Verfahren liefert offenbar auch bei der relativen Einordnung der Resultate beider Suchmustermengen soweit stabile Resultate, dass hinsichtlich des

[18] Eine Wiederverwendung der bereits annotierten Stichproben – für die wir uns im Falle der Präzisionsschätzungen für die automatisch erzeugten Suchmuster entschieden haben – wäre hier nicht sinnvoll, da es sich bei diesen sämtlich um Treffer handelt, die bereits von den manuell spezifizierten Suchmustern extrahiert werden.

**Gesamtergebnisse**

| *p* | *r* | *f* |
|---|---|---|
| 0,16 | 0,62 | 0,25 |

**Ranking**

| **Max. f-Score** | **MAP** | **Recall Level** | | | | |
|---|---|---|---|---|---|---|
| | | *0* | *0,1* | *0,2* | *0,3* | *0,4* |
| 0,40 (p=0,46, r=0,35) | 0,29 | 1 | 0,68 | 0,52 | 0,46 | 0,33 |

Tabelle 6.12: Ergebnisse der Definitionssuche mit manuell und durch Bootstrapping erzeugten Suchmustern

maximalen f-Score (der mit 0,4 den ursprünglich für die Entwicklungsdaten erzielten Wert erreicht) und der MAP signifikante Verbesserungen gegenüber den besten Ergebnissen auf der Basis der manuell spezifizierten Suchmuster zu Stande kommen.[19] Auch die punktbezogenen Präzisionswerte liegen durchgängig im Bereich der besten bisher erzielten Scores oder noch darüber.

### 6.3.3 Analyse und Bewertung

Das diskutierte Experiment hat gezeigt, dass sich schon unter Verwendung eines relativ einfachen Bootstrapping-Verfahrens Suchmuster von vergleichsweise hoher Qualität erzeugen lassen. Die durch Bootstrapping gewonnenen Suchmuster sind zu den manuell erzeugten soweit komplementär, dass sich durch die Kombination beider Musterbestände signifikant bessere Ergebnisse erzielen lassen als durch die einzelnen Suchmustermengen. Der unter dem optimalen Ranking der kombinierten Suchmuster maximal erreichte f-Score für die Extraktion aus unserem Goldstandard-Korpus beträgt 0,4. Dies entspricht dem Wert, der durch die nicht-optimierten manuell erzeugten Suchmuster auf den bei der Spezifikation genutzten Entwicklungsdaten erreicht wurde.

Auch die Ergebnisqualität der automatisch erzeugten Suchmuster alleine liegt bei Verwendung der in diesem Kapitel entwickelten Ranking-Methoden

[19] Die Signifikanz der MAP-Differenz wurde wiederum durch einen Randomisierungstest bestätigt.

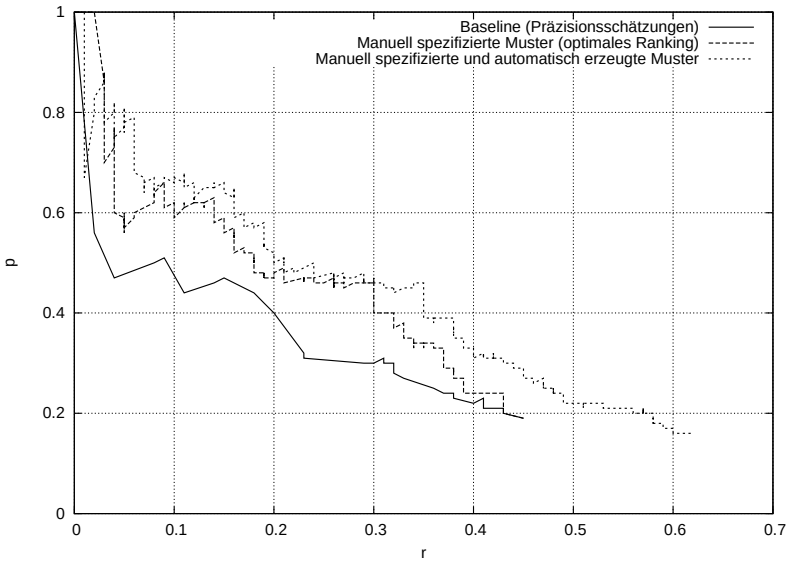



Abbildung 6.7: Ranking der kombinierten Extraktionsergebnisse (Bootstrapping und manuell spezifizierte Suchmuster)

in einem ähnlichen Bereich wie die der manuell spezifizierten Muster. Die Inspektion der durch den Bootstrapping-Prozess erzeugten Extraktoren lässt allerdings vermuten, dass deren Performanz bei der hier nicht untersuchten Aufgabenstellung der Identifikation von Definitionsbestandteilen (vgl. 5.3.3 und 5.3.3) deutlich unter diejenige der manuell spezifizierten Muster zurückfallen dürfte, da ein Teil der automatisch erzeugten Suchmuster unplausible Pfade zu den Definitionsbestandteilen enthält.

Aufgrund unserer verhältnismäßig einfachen Umsetzung des Bootstrapping-Zyklus ist mit weiterem Verbesserungspotential durch Optimierungen in den einzelnen Schritten zu rechnen. Insbesondere betrifft dies die Auswahl der in jedem Zyklus beizubehaltenden Suchmusterkandidaten (bei deren Ranking eine Berücksichtigung des Zugewinns an bisher noch nicht extrahierbaren Treffern erfolgen sollte) und der neu erzeugten Seeds (die etwa im Hinblick auf ihre Relation zu bereits verwendeten Seeds sowie allgemeine Frequenzdaten bewertet

werden sollten). Möglicherweise kann das Verfahren durch solche Änderungen soweit optimiert werden, dass zusätzliche Durchläufe zu einer weiteren Ergebnisverbesserung führen.

## 6.4 Fazit

Wir haben in diesem Kapitel verschiedene Methoden zur Optimierung der Suchergebnisse unseres Definitionsextraktionssystems erprobt und miteinander verglichen. Zunächst haben wir den Ansatz verfolgt, Präzision und Recall durch punktuelle, wissensintensive Maßnahmen zu optimieren, die direkt an Schwachpunkten unseres Extraktionssystems ansetzen, die wir im vorigen Kapitel ermittelt haben. Zur Verbesserung der Präzision haben wir versucht, mit verschiedenen Verfahren auf der Basis des annotierten Entwicklungskorpus eine optimale Untermenge unseres Suchmusterbestandes zu erzeugen sowie durch eine Stoppwort-basierte Filterung möglichst viele *false positives* aus den Ergebnissen zu entfernen. Zur Recall-Verbesserung haben wir (ausgehend von Korpusfrequenzen) zusätzliche Suchmuster für einige besonders häufig auftretende Definitionsprädikate spezifiziert, die durch den bisherigen Musterbestand nicht abgedeckt waren.

In beiden Fällen waren die erzielten Veränderungen jedoch nicht zufriedenstellend. Insbesondere waren sie nicht hinreichend, um den teils erheblichen manuellen Aufwand zu rechtfertigen, der mit den eingesetzten Verfahren verbunden ist. Im zweiten Teil des Kapitels haben wir daher in höherem Maße automatisierte Methoden untersucht, die gleichzeitig stärker datenorientiert arbeiten und eine größere Vielfalt an Informationsquellen einbinden. Zur Präzisionsverbesserung haben wir verschiedene klassifikationsbasierte Filter erprobt, für die zu den Extraktionsergebnissen ein wesentlicher Teil der in Kap. 2 diskutierten “weichen” Definitionskennzeichen in Merkmalsvektoren kodiert wurde. Optimale Resultate konnten wird jedoch nicht durch eine solche Filterung, sondern durch ein auf dieselben Merkmale gestütztes Ergebnisranking erzielen.

Eine Recall-Verbesserung – die bei Verwendung eines entsprechenden Ergebnisranking auch zu einer deutlichen Verbesserung der Gesamtergebnisse führt – erbrachte die automatische Erzeugung neuer Suchmuster durch ein Bootstrapping-Verfahren. Insgesamt konnte so eine Verbesserung des f-Score (gemessen für die optimale Bestenliste des Ranking) um über ein Drittel gegenüber dem besten in 6.1 erzielten Ergebnis erreicht werden.

Das verwendete Bootstrapping-Verfahren lässt an mehreren Stellen offensichtliche Optimierungen zu, die eine weitere Verbesserung der Gesamtergebnisse möglich erscheinen lassen. Auch die Effektivität des Ranking-Verfahrens kann mit großer Wahrscheinlichkeit durch Veränderungen an den verwende-

ten Merkmalen und eine aufwändigere Datenaufbereitung (etwa die Erstellung separater Trainingskorpora für Definitionen mit Kopula und Oberbegriff und solche mit anderen Prädikaten, vgl. Kap. 2) noch verbessert werden. Die erzielbare Ergebnisqualität bleibt andererseits in jedem Fall durch die Probleme der linguistischen Vorverarbeitung begrenzt (siehe 5.4).

Im Rahmen der hier beschriebenen Arbeiten konnte kein Anwendungsszenario vertieft untersucht werden (vgl. zu den in Form von *proofs of concept* betrachteten Anwendungen die kurze Diskussion im nächsten Kapitel, 7.2). Wir sind jedoch überzeugt, dass die meisten solchen Szenarios weitere, über die hier diskutierten hinausgehende Optimierungsschritte ermöglichen werden. So wäre es zum Beispiel im Rahmen einer Suchmaschine zum stichwortbasierten Auffinden von Definitionen in einer Entscheidungssammlung sinnvoll, in dem von uns entwickelten Ranking-Verfahren auch suchbegriffspezifische Information einzubeziehen. Damit kann nicht nur eine hohe Relevanz der in das Top-Segment sortierten Treffer für den Suchbegriff sichergestellt werden, sondern – etwa durch den Vergleich mit einem Begriffsprofil, vgl. 3.2.3 – auch dessen "Definitionsdichte" erhöht werden. Durch die Erfassung von Benutzer-Feedback lässt sich in Kombination mit dem beschriebenen Bootstrapping-Verfahren in einem kontinuierlichen Qualitätssicherungsprozess der Suchmusterbestand ständig weiter optimieren und aktuell halten. Schließlich kann in einem solchen Rahmen – unter Verwendung einer entsprechend erweiterten Ranking-Komponente – auch das in 5.1 diskutierte Lemma-basierte Extraktionsverfahren als Fallback-System zur Steigerung des Recall integriert werden, um den negativen Einfluss der Beschränkungen der linguistischen Vorverarbeitung aufzufangen.

# Kapitel 7

# Ausblick und Schluss

In dieser Arbeit haben wir das Thema “juristische Definitionen” aus zwei sehr verschiedenen Perspektiven beleuchtet: Der erste Teil der Arbeit enthält eine sprachtheoretische Untersuchung, bei der wir uns bemüht haben, die fachspezifische Dimension des Phänomens Definition in der Rechtssprache mit ins Auge zu fassen. Der zweite Teil untersucht dagegen die computerlinguistische Fragestellung, wie solche Definitionen mit sprachtechnologischen Mitteln identifiziert und verarbeitet werden können.

Beide Teile sind durch den verfolgten empirisch-korpusbasierten Ansatz verbunden. Die im ersten Teil erarbeitete sprachwissenschaftliche Beschreibung juristischer Definitionen diente darüberhinaus als Grundlage für die Entwicklung des im zweiten Teil diskutierten musterbasierten Extraktionsverfahrens. Das Textkorpus, das wir im Rahmen der linguistischen Untersuchung aufbereitet haben, lieferte die Datengrundlage für die Erstellung und Validierung des implementierten Systems.

Es verbleiben jedoch noch verschiedene weitere Anknüpfungspunkte zwischen den beiden Teilen, denen wir im Rahmen der vorliegenden Arbeit nicht nachgehen konnten. Diese dürften insbesondere vor dem Hintergrund denkbarer Anwendungen des Definitionsextraktionsverfahrens relevant werden. Wir geben in diesem Kapitel zunächst eine zusammenfassenden Gesamtbewertung des Erreichten. Dann gehen wir auf zwei wichtige direkte Anwendungsperspektiven unserer Ergebnisse ein und stellen kurz einige Vorarbeiten dar, die wir in diesen Bereichen geleistet haben.

## 7.1 Ergebnisse der Arbeit

In dieser Arbeit haben wir eine Methode zur automatischen Identifikation von Definitionen in Gerichtsurteilen entwickelt und bis zur Praxisreife ausgearbeitet. Wir haben dabei in hohem Maße domänenspezifische Besonderheiten berücksichtigt.

Im interdisziplinären Bereich zwischen Computerlinguistik und Rechtswissenschaft gab es bisher keine vergleichbaren Forschungen. Wir haben im Rah-

men dieser Arbeit daher an vielen Stellen Neuland beschritten. Auch wenn die gewonnenen Erkenntnisse somit an einzelnen Stellen noch weiterer Vertiefung und Validierung bedürfen, haben wir auf theoretischer und auf praktischer Ebene relevante Beiträge geleistet:

In linguistisch-theoretischer Hinsicht besteht der Hauptbeitrag unserer Arbeit darin, dass das Phänomen rechtssprachliche Definition erstmals auf einer großen Korpusgrundlage empirisch untersucht wurde. Die so gewonnenen Einsichten dürften – auch wenn dieser Fragestellung hier nicht explizit nachgegangen wurde – zu einem beträchtlichen Teil auch auf Definitionen in anderen Textsorten übertragbar sein. Hinsichtlich der Rolle von Definitionen in Rechtstexten konnten die bisherigen, stark deduktiv-wissenschaftstheoretisch geprägten rechtstheoretischen Überlegungen um eine komplementäre datenbasierte Ebene ergänzt werden. Die Funktion richterlicher Definitionen in der juristischen Argumentation wurde so in einer empirischen Breite und Differenziertheit beschrieben, die bisher in der Auslegungslehre nicht vorhandenen war. Besondere praktische Bedeutung erhalten die Ergebnisse dadurch, dass die erarbeiteten Kategorisierungen – anders als in den meisten rechtstheoretischen Grundlagenwerken – direkt durch eine Vielzahl von Belegstellen auf konkrete richterliche Argumentationen in realen Fällen bezogen sind.

Im praktisch-technischen Bereich liegt die Innovativität der Arbeit vor allen Dingen in der untersuchten Aufgabenstellung und Anwendungsdomäne. Sowohl die Extraktion von Definitionen als eigenständige Aufgabe (also nicht als Teilaufgabe etwa in einem *Question Answering*-System) als auch die Anwendung von Verfahren des textbasierten Informationszugriffs auf Rechtstexte sind Forschungsbereiche, die sich gegenwärtig erst formieren. So wurde 2009 erstmals ein internationaler Workshop zum Thema Definitionsextraktion veranstaltet und 2006 wurde zum ersten Mal ein TREC-Track zum juristischen Information Retrieval abgehalten.

Wir fassen nun die einzelnen Ergebnisse in diesen beiden Bereichen noch einmal zusammen:

**Analyseschema für richterliche Definitionen auf formaler, semantischer und pragmatischer Ebene:** Wir haben die verschiedenen Formen und Funktionen von Definitionen in Urteilsbegründungen umfassend systematisiert. Unser Analyseschema erfasst den Zusammenhang zwischen sprachlicher Form, semantischem Gehalt und juristischer Funktion solcher Äußerungen. Wir haben dabei auf Vorarbeiten aus Sprachphilosophie und Semantik, Rechtstheorie sowie aus der Terminologiewissenschaft aufgebaut, über die wir aber insbesondere in folgenden Punkten wesentlich hinausgehen:

- Wir haben einen rein deskriptiven Ansatz verfolgt und uns durchgängig auf reichhaltige empirische Evidenz in Form von Korpusbeispielen gestützt.
- Wir haben eine breite Palette unvollständiger Definitionstypen in unsere Untersuchung einbezogen, die von den traditionell hauptsächlich betrachteten taxonomischen Definitionsschemata abweichen.
- Unsere Analyse erfasst außerdem auch den Aspekt der textstrukturierenden und argumentativen Funktion von Definitionen, der im Rahmen der Textsorte *Urteilsbegründung* eine besonders wichtige Rolle spielt.

Insgesamt hoffen wir, damit ein Stück zu einem genaueren Verständnis des Prozesses der Konstruktion und Adaption von Bedeutung mit sprachlichen Mitteln beigetragen zu haben.

**Sorgfältig annotiertes Korpus von Urteilsbegründungen:** Ein Produkt dieser Analysen ist ein kontrolliert im Hinblick auf Definitionen annotiertes Korpus. Die Annotationsrichtlinien sind in Walter (2006) ausführlich dokumentiert. Für einen wesentlichen Teil des Korpus (60 der 100 analysierten Entscheidungen) liegen doppelte Annotationen vor. Diese wurden auf Übereinstimmung analysiert und in einem Konsensverfahren zu einem Goldstandard zusammengeführt. Bei dem erzeugten Datenbestand handelt es sich unseres Wissens nicht nur um das einzige Korpus der deutschen Rechtssprache, das in dieser Form aufbereitet ist. Auch unabhängig von Sprache und Domäne sind uns keine vergleichbaren Ressourcen bekannt.

**Inventar von Definitionsformulierungen:** Auf der Grundlage dieses Korpus konnten wir ein umfassendes Inventar definitorischer Formulierungsmuster der deutschen Rechtssprache zusammenstellen, das wir unter sprachlich-formalen Gesichtspunkten analysiert und systematisiert haben. Zu unterscheiden ist hier auf oberster Ebene zwischen parenthetischen und prädikatbasierten, also in ganzen Sätzen realisierten Definitionen. Diese lassen sich weiterhin nach verwendetem Prädikat und den Mitteln zum Anschluss der Definitionsbestandteile einteilen. Die Formulierungsmuster für prädikatbasierte Definitionen dienten als Ausgangspunkt zur Erstellung der Suchmuster für ein regelbasiertes Definitionsextraktionsverfahren.

**Sprachtechnologische Werkzeuge reicher Textrepräsentationen für Rechtstexte:** Aufbauend auf dem PreDS-Parser aus Fliedner (2007) als Kernkomponente haben wir eine vollständige Verarbeitungskette zur Erzeugung reicher

Textrepräsentationen für juristische Dokumente aufgebaut. Der PreDS-Parser, der in verschiedenen vorangegangenen Projekten entwickelt und erprobt wurde, erzeugt robust tiefe linguistische Analysen für deutschen Text. Wir haben den Parser weiterentwickelt, um den Besonderheiten der Rechtssprache gerecht werden zu können. Unsere Erweiterungen umfassen optimierte Regeln für die Verarbeitung typisch rechtssprachlicher "Problemfälle", wie z.B. tief eingebetteter Nominalphrasen und Relativsatzschachtelungen. Darüberhinaus beinhalten sie eine spezialisierte Komponente zur Erkennung und korrekten Integration juristischer Zitate in die Satzstruktur mit breiter Abdeckung. Die gesamte Verarbeitungskette umfasst neben dem PreDS-Parser auch Module für die Textnormalisierung und -strukturierung sowie eine Komponente zur Korpusverwaltung (Ablage und Abfrage der erzeugten linguistischen Analysen). Der PreDS-basierte tiefe Analyseweg wird durch Fallback-Komponenten auf der Basis von Standardwerkzeugen zur flachen linguistischen Analyse (Lemmatisierung und POS-Tagging) ergänzt.

**Großes automatisch analysiertes Korpus und statistische Ergebnisse zu rechtssprachlichen Besonderheiten:** Mittels der angesprochenen Verarbeitungskette haben wir insgesamt mehr als 33 000 Entscheidungstexte (die uns die Firma *juris* freundlicherweise verfügbar gemacht hat) automatisch mit umfassender linguistischer Strukturinformation versehen. Dieser Textbestand diente uns als Datengrundlage für die Entwicklung, Erprobung und Optimierung unseres Definitionsextraktionsverfahrens. Das Korpus ermöglicht jedoch aufgrund seines Umfangs auch eine neue Perspektive auf Fragestellungen der juristischen Fachsprachenforschung. Wir haben auf seiner Basis verschiedene häufig beschriebene Beobachtungen über die Eigenheiten und die Komplexität der deutschen Rechtssprache empirisch überprüft (und weitestgehend bestätigen können), die bisher meist nur durch Einzelbeispiele belegt waren.

**Extraktionsverfahren:** Hauptergebnis der Arbeit ist jedoch ein Verfahren zur automatischen Extraktion und Verarbeitung von Definitionen in Urteilstexten. Auf der Grundlage dieses Verfahrens habe wir das CORTE-System implementiert, in dem Definitionen aus Dokumenten extrahiert werden, die die oben beschriebene Verarbeitungskette durchlaufen haben.

Kern des Extraktionsverfahrens ist ein regelbasierter Suchschritt. Dessen Suchmusterbestand setzt direkt das Inventar definitorischer Formulierungen um, das wir bei unserer Korpusanalyse identifiziert haben. Eine vergleichende Evaluation zeigt, dass gegenüber einer Suche auf der Basis flacher Analysen ein erheblicher Präzisionszugewinn durch die Hinzunahme linguistischer Strukturinformation aus den PreDS-Repräsentationen erzielt wird. Schon ohne

weitere Optimierungen wird so eine Ergebnisqualität erreicht, die den Einsatz des CORTE-Systems für Anwendungen mit manuellen Nachbearbeitungsmöglichkeiten erlaubt.

Durch den Einsatz datengesteuerter Optimierungen konnten wir die Leistungsfähigkeit unseres Definitionsextraktionsverfahrens noch erheblich erhöhen. Die besten Resultate werden durch die Nutzung eines statistisch arbeitenden Ranking-Verfahrens in Kombination mit einem Bootstrapping-Ansatz zum Erwerb zusätzlicher Suchmuster erzielt. Die so erreichten Verbesserungen erweitern den möglichen Einsatzbereich des CORTE-Systems. Um dies unter Beweis zu stellen, haben wir es unter anderem in einen Suchmaschinen-Prototypen zum stichwortbasierten Zugriff auf Definitionen in einer Entscheidungssammlung integriert. Zudem haben wir die Verwendung der Ergebnisse der Definitionsextraktion für die Optimierung einer Methode zur automatischen Identifikation juristischer Terminologie erprobt. Beide Experimente lieferten positive Ergebnisse (vgl. 7.2).

Wir konnten somit zeigen, dass für Text Mining-Aufgaben im juristischen Bereich wissensbasierte Verfahren, die sich auf eine sorgfältige konzeptuelle Domänenmodellierung und tiefe linguistische Information stützen, einen erfolgversprechenden Ansatzpunkt bieten. Zumindest teilweise dürfte Ähnliches auch für andere abgeschlossene Domänen mit geringen Redundanzen gelten (etwa für Sammlungen historischer Texte oder spezieller Fachliteratur). Ein mit begrenztem Expertenaufwand erstelltes wissensbasiertes System kann dann durch Hinzunahme flacher statistischer Verfahren halbautomatisch verbessert und erweitert werden, um beispielsweise eine erhöhte Abdeckung zu erzielen und Engpässe beim Wissenserwerb zu kompensieren.

## 7.2 Anwendungsperspektiven

Für die praktische Nutzung der in dieser Arbeit erzielten Ergebnisse und Erkenntnisse sehen wir mindestens zwei direkte Perspektiven:

(a) Die Ergänzung eines juristischen Informationssystems um eine Definitionssuche bzw. die Aufbereitung der Suchergebnisse in einem solchen System (wir haben diese Einsatzmöglichkeit im Rahmen der Arbeit mehrfach angesprochen). Der Nutzerkreis besteht in diesem Fall aus den Nutzern des Informationssystems, also Juristen (oder ratsuchenden juristischen Laien).

(b) Den Einsatz im juristischen *knowledge engineering*, also bei der strukturierten, u.U. auch formalen, Modellierung juristischen Wissens. Nutzer-

kreis sind hier zunächst nur die an einem solchen Modellierungsprozess beteiligten Personen, bei denen es sich meistens um theoretisch arbeitende Juristen, aber auch um Terminologieexperten oder informationstechnologische Fachleute handeln wird.

### 7.2.1 Definitionssuche in einem juristischen Informationssystem

In einem juristischen Informationssystem kann eine Definitionsextraktionskomponente beispielsweise bei der Indexerstellung (also zur automatischen Erzeugung bzw. Pflege von Ressourcen wie dem *juris*-Definitionsindex, vgl. 2.4.2) oder bei der Relevanzgewichtung von Suchergebnissen genutzt werden (Treffer mit einer Definition des Suchbegriffs sind mit besonders hoher Wahrscheinlichkeit relevant). Der naheliegendste Einsatzbereich ist jedoch eine Suchfunktion zum direkten Zugriff auf Definitionen in einer Entscheidungssammlung. Diese kann zusätzlich mit einer Aufbereitung oder genaueren Strukturierung der Ergebnisse gekoppelt werden.

Wir haben die in unserer Arbeit vorgestellten Ergebnisse in einem prototypischen System zur gezielten stichwortbasierten Suche nach Definitionen zusammengeführt. Implementiert wurde eine Suchmaschine mit einer browserbasierten Benutzeroberfläche. Sie liefert zu Benutzereingaben in Form von Suchbegriffen Fundstellen zurück, die nach Treffern in Definiendum, Definiens und Volltext gruppiert sind. Für diese Einteilung der Fundstellen wird auf die automatisch ermittelte Definitionssegmentierung zurückgegriffen. Die Definitionsbestandteile werden in den Suchergebnissen zudem typographisch hervorgehoben. Die Trefferlisten sind jeweils nach dem in 6.2.4 beschriebenen kombinierten Rankingverfahren sortiert. Sie können auf Benutzeranfrage dynamisch um den Dokumentkontext sowie Informationen zu juristisch relevanten Metadaten erweitert werden. Als Datengrundlage diente das gesamte von *juris* zur Verfügung gestellte Korpus.

Eine Auswertung der implementierten Suchmaschine müsste neben der Ergebnisqualität auch die Benutzbarkeit und Effektivität der Ergebnispräsentation berücksichtigen. Beide Fragen erfordern eine intensive Kooperation mit juristischen Experten und konnten im hier diskutierten Rahmen nicht mehr angegangen werden. Die durchgeführten Tests haben jedoch verschiedene Stellen aufgezeigt, an denen mit großer Wahrscheinlichkeit Verbesserungen auf der Grundlage von Ergebnissen und Erkenntnissen aus dieser Arbeit erreicht werden können. Hierzu gehören zum Beispiel folgende Ansatzpunkte:

- Anhand der in Kap. 2 ausgearbeiteten Typisierung kontextualisierter und argumentativ gebrauchter Definitionen könnte eine zusätzliche Sortie-

rung der Ergebnisse nach Definitionstypen und "definitorischem Wert" erfolgen. Zumindest dürfte eine Einteilung von Suchergebnissen in Legaldefinitionen, Spezialisierungen auf Fallklassen (wie *Gebäude mit Fundament* in der im ersten Kapitel ausführlich untersuchten Argumentation) und subsumtionsnahe fallspezifische Festlegungen sinnvoll sein. Hierzu wäre zunächst der in Kap. 2 nur in ersten Schritten entwickelte Ansatz der Identifikation und Zuordnung von Schlüsselwörtern (wie Modifikatoren und den in 2.3.2 angesprochenen *Hedges*) weiter auszubauen.

- Mit Hilfe der Definitionssegmentierung oder auch sprachlicher Hinweise auf die diskursive Einbettung könnten Suchergebnisse entlang der Achsen *Begriffsbeziehungen* bzw. *Argumentationszusammenhang* gruppiert und um weitere Fundstellen ergänzt werden. So wie beispielsweise in den Ergebnislisten der *juris*-Suche Zitatangaben über Hyperlinks verfolgt werden können, könnten in Ergebnissen unserer Definitionssuche Definitionen der verwendeten Genusbegriffe sowie übereinstimmende und konkurrierende Definitionen aus anderen Urteilen verlinkt werden.

- Die vorliegenden linguistisch aufgearbeiteten Korpora können nicht nur als durchsuchbare Datengrundlage eingesetzt werden, sondern zusätzlich als Hintergrundressource etwa für die Ermittlung ähnlicher oder korrelierter Begriffe zu Benutzeranfragen. Diese können z.B. für *query expansions* oder zur Erzeugung von Suchfacetten genutzt werden (so könnten beispielsweise die Suchvorgänge *Abfälle zur Beseitigung* und *Abfälle zur Verwertung* für den Suchbegriff *Abfall* vorgeschlagen werden).

- Durch die Erfassung und Auswertung von Benutzerfeedback könnten in einem kontinuierlichen Qualitätssicherungs- und Verbesserungsprozess der Suchmusterbestand aktuell gehalten, die verwendete Ranking-Komponente weiter trainiert und der Musterauswahlschritt des in 6.3 beschriebenen Bootstrappingverfahrens optimiert werden.

Unser Suchmaschinenprototyp stellt den Rahmen für eine zielgerichtete Erprobung dieser Verbesserungsansätze zur Verfügung.

### 7.2.2 Induktive Wissensgewinnung

Die Tatsache, dass ein Konzept in einem Text definiert wird, spricht dafür, dass es für Thema oder Funktion des Textes besonders relevant ist. Im Falle von juristischem Text erhält diese Annahme noch zusätzliche Plausibilität. Juristische Argumentationen stützen sich nämlich in der Regel – wie in den ersten beiden

Kapiteln dieser Arbeit ausführlicher erläutert – an den entscheidenden Punkten auf die definitorische Ausarbeitung der für den Fall wesentlichen Rechts- und Domänenbegriffe. Definitionsextraktion ist daher ein vielversprechender Ausgangspunkt zur Identifikation wesentlicher konzeptueller Strukturen in der Rechtsdomäne.

Traditionell wird solches Wissen in rechtstheoretischen Werken und juristischen Kommentaren zusammengetragen und ausgearbeitet. In jüngerer Zeit formiert sich zudem der Themenbereich *juristische Ontologien* als eigenes Forschungsfeld, in dem Rechtsbegriffe und ihre Zusammenhänge mit formalen Mitteln erfasst und computerlesbar kodiert werden. Dass die Ergebnisse einer automatischen Definitionssuche direkte Verwendung finden könnten, ist in beiden Fällen offensichtlich. Während allerdings die Arbeitsweise eines Rechtswissenschaftlers, beispielsweise bei der Kommentarerstellung, vermutlich gut auf der Grundlage einer erweiterten Suchmaschine unterstützt werden kann (vgl. die Überlegungen im letzten Abschnitt), bietet die Ontologieerzeugung und -pflege erheblich höheres Automatisierungspotential.

Zu diesem Zweck müssten, ausgehend von der in Kap. 5 entwickelten Definitionssegmentierung, formale Repräsentationen extrahierter textueller Definitionen erzeugt werden. Dabei wären insbesondere die in Kap. 2 diskutierten unterschiedlichen Status von Definitionen in Urteilstexten zu berücksichtigen. Die erzeugten Repräsentationen könnten dann mit logischen Verfahren konsolidiert und in bestehende Ontologien eingepflegt werden.

Logisch zu modellieren wäre beispielsweise für die Definition in 7.1 (vereinfacht dargestellt) der Effekt, dass die Hyperonymierelation zwischen “öffentlich-rechtlicher Vertrag” und “Vertrag” hinzugefügt und die vom Oberbegriff geerbte Rolle des Vertragsgegenstandes auf den öffentlich-rechtlichen Sachbereich eingeschränkt wird.

(7.1) Öffentlich-rechtlich sind diejenigen Verträge, deren Gegenstand einem vom öffentlichen Recht geordneten Sachbereich zuzuordnen ist.

(BayObLG München 5. Zivilsenat, 16. Juli 2001, 5Z RR 73/98, juris)

Besonders geeignet erscheinen hierfür beschreibungslogische Formalismen, in denen das Update einer Wissensbasis um solche Information als monotone Erweiterung eines Gesamtmodells dargestellt werden kann.

Die Unterstützung solcher Prozesse auf der Basis unseres Definitionsextraktionssystems setzt zunächst eine weitere Erhöhung der Ergebnisqualität voraus (insbesondere im Hinblick auf die Definitionssegmentierung). Viele der beteiligten semantischen Vorgänge (etwa die in Kap. 1 angesprochenen Präzisierungsmechanismen des rechtssprachlichen *open texture*) sind zudem noch nicht

im Detail geklärt und modelliert. Auch für die automatische Erzeugung beschreibungslogischer Bedeutungsrepräsentationen mit computerlinguistischen Techniken wurden noch keine Verfahren voll ausgearbeitet. Schließlich existieren Rechtsontologien bisher nur in Form kleiner Prototypen und es ist noch kein allgemein anerkanntes *upper level*-Modell der Rechtsdomäne verfügbar. Im Bereich der automatischen Induktion juristischen Wissens ist also in unmittelbarer Zukunft nicht mit Ergebnissen zu rechnen, die über Machbarkeitsstudien und exemplarische Lösungen von Einzelproblemen hinausgehen.

Wir haben als *proof of concept* für dieses Einsatzfeld die Verwendbarkeit unseres implementierten Definitionsextraktionssystems zur Verbesserung der Ergebnisse einer traditionellen Terminologieextraktionsmethode untersucht. Diese Aufgabenstellung beeinhaltet zum einen bereits an sich ein erhebliches Nutzpotential. Zum anderen kann sie als "leichtere Variante" des Ontologieupdate angesehen werden. Der bis dato auch theoretisch noch nicht vollständig geklärte Aspekt der automatischen Erzeugung formaler, logisch verarbeitbarer Bedeutungsrepräsentationen aus Textextrakten bleibt dabei außen vor.

Ein gängiges allgemeines Verfahren zur korpusbasierten Identifikation von Terminologie ist die Extraktion von stark korrelierten Adjektiv-Substantiv-Bigrammen. So extrahierte Termini können zugleich einen Ansatzpunkt für den Aufbau von Ontologiefragmenten bieten: Im Normalfall impliziert das durch das Bigram bezeichnete Konzept den durch das Substantiv bezeichneten Begriff, im Idealfall handelt es sich bei dieser Beziehung um eine auch ontologisch relevante konzeptuelle Unterordnung. Wir haben das CORTE-System für die Filterung von Terminologie-Kandidaten verwendet, die mittels eines *log-likelihood*-Rankings von Adjektiv-Substantiv-Bigrammen im umweltrechtsbezogenen Teil des CORTE-Korpus identifiziert wurden. Eine Filterung der extrahierten Bigramme nach Auftreten in automatisch identifizierten Definitionen führte zwar zu einer starken Präzisionserhöhung, jedoch auch zu einer drastischen Reduktion der Abdeckung. Ein optimales Gesamtergebnis konnten wir durch eine Kombination des Top-Segments der ungefilterten mit den Ergebnissen der gefilterten Extraktion erzielen. Mit diesem kombinierten Verfahren konnte die Anzahl der bei brauchbarer Präzision (0.5 und mehr) extrahierten Termkandidaten verfünffacht werden.

Der erprobte Ansatz kann in vieler Hinsicht verfeinert werden. In unseren Versuchen haben wir weder von den Ergebnissen der Definitionssegmentierung (und somit der Möglichkeit, gezielt Definienda zu identifizieren) Gebrauch gemacht, noch von den Optimierungen, die wir in Kap. 6 beschrieben haben. Verbesserungen können außerdem sicherlich durch den Einsatz einer weiter entwickelten Baseline-Termextraktionsmethode und durch die Hinzunahme weiterer Komponenten, z.B. einer Ähnlichkeitssuche anhand bekannter Terminologie, erzielt werden. Direkt einsetzbar ist das Verfahren in der beschriebenen

Form wohl bereits zur Unterstützung von Experten etwa bei der Dokumentation und Verschlagwortung von Entscheidungen.

Insgesamt haben bereits diese ersten Experimente deutlich gezeigt, dass mit dem von uns entwickelten Definitionsextraktionsverfahren eine vielfältig einsetzbare Methode zur Erschließung des konzeptuellen Wissens in Entscheidungstexten zur Verfügung steht. Über unsere wissenschaftlichen Ergebnisse und die praktischen Perspektiven hinaus wäre es jedoch schön, wenn diese Arbeit noch einen weiteren, interdisziplinären Beitrag leisten könnte: einen Ausgangspunkt für neue Kooperationen zwischen Computerlinguisten und Rechtswissenschaftlern zu bieten.

# Anhang A: Verwendete Korpora

| *Bezeichnung* | *Beschreibung* | *Annotierte Information* | *Nutzung* |
|---|---|---|---|
| *Pilotstudien-Korpus* (siehe 2.1) | Entscheidungstexte (40 Dokumente, 3757 Sätze 127 349 Wörter) | Definitionen; parenthetisch vs. prädikatbasiert; Definitionsprädikate; Informationsschicht (Kern vs. Elaboration); argumentative Funktion; Definitionsbestandteile; Automatisch ermittelt: POS / Lemmata, PreDS | Entwicklungsdaten für Systematisierung von Definitionen; Entwicklungsdaten für Erstellung der Suchmuster |
| *Goldstandard-Korpus* (siehe 2.4) | Entscheidungstexte (60 Dokumente, 7627 Sätze 233 210 Wörter) | Definitionen (doppelt, $\kappa = 0,58$); parenthetisch vs. prädikatbasiert; Definitionsprädikate; Informationsschicht (Kern vs. Elaboration, doppelt, $\kappa = 0,56$); Definitionsbestandteile; Automatisch ermittelt: POS / Lemmata, PreDS | Definitionsannotation; Testdaten für Definitionsextraktion und -segmentierung |
| *Juris-(Groß)korpus* (siehe 4.3) | Entscheidungstexte (33 597 Dokumente, 2 307 189 Sätze 71 409 021 Wörter) | Trefferauswertung: 6000 dependenzbasiert 6000 sequenzbasiert (je 2000 doppelt, $\kappa = 0,83$ bzw. $\kappa = 0.85$) Automatisch ermittelt: POS / Lemmata, PreDS | Testdaten für Definitionsextraktion; Trainingsdaten für Klassifikation und Ranking; Datengrundlage für Anpassung des Preds-Parsers; statistische Untersuchung rechtssprachlicher Besonderheiten |
| *ECI-FR-Korpus* (siehe 4.3.2) | Zeitungstext (40 856 149 Wörter) | Automatisch ermittelt: POS / Lemmata (Gesamtkorpus); PreDS (TIGER-Korpus, 900 000 Wörter) | Vergleichsdaten (Untersuchung rechtssprachlicher Besonderheiten) |

# Anhang B: Systeme für Definitions-QA

**TREC**

| *System (Rank)*[1] | *Muster / Analysen* | *Zusatzinformation / Besonderheiten* |
|---|---|---|
| *BBN (1)* [Xu u. a. (2003, 2004)] | 40 bzw. 22 / Wortfolgen, Präd.-Arg.-Tripel | Profile (Begriff, Definition und Definitionstyp) aus Treffern, Online-Enzyklopädien, Biographien; vordefiniertes Ranking der Muster / ROUGE für automatische Evaluation; chin. Version in Peng u. a. (2005) |
| *Qualifier (2)* [Yang u. a. (2003)] | k.A. / Wortfolgen | Begriffsprofile (aus dem WWW) / verwenden Sätze ohne den Suchbegriff für "Negativprofil" |
| *TextMap (3)* [Echihabi u. a. (2003)] | 110 / Semantische Relationen und Begriffsprofile (biographische Deskriptoren, WWW, Wordnet) | |
| *LCC QA (4)* [Harabagiu u. a. (2003)] | 38 / Wortfolgen | |
| *DefScriber (5)* [Blair-Goldensohn u. a. (2003a,c,b)] | 18 / Syntax | Definitionsprofile (aus allen Treffern); verschiedene vordefinierte Merkmale / kohärenter Text als Antwort |
| *UAmsterdam (7)* [Jijkoun u. a. (2003); Ahn u. a. (2004)] | keine / k.A. | Begriffsprofile (WWW, Wikipedia, Wordnet) / Extraktion mittels Begriffsprofilen |
| *MIT CSAIL (8)* [Katz u. a. (2003); Hildebrandt u. a. (2004)] | 13 / Chunks / Begriffsprofile (Merriam-Webster online) | |
| *USheffield (9)* [Gaizauskas u. a. (2003); Saggion (2004)] | 50 / Wortfolgen / Begriffsprofile (Online-Enzyklopädie, WWW, Wordnet) | |
| *IBM PIQUANT (14)* [Prager u. a. (2003)] | k.A. / Syntax | Wordnet / Erzeugung eines "Dossiers" aus Faktoid-Fragen (2003) |

**TREC**

| *System (Rank)*[1] | *Muster / Analysen* | *Zusatzinformation / Besonderheiten* |
|---|---|---|
| *IBM*<br>[Prager u. a. (2001)] | keine /<br>Lemmata | Wordnet /<br>TREC 9; Extraktion durch Suche nach Wordnet-Hypernymen; Evaluation der Top 5-Akkuratheit |
| *UKorea*<br>[Han u. a. (2005, 2006a,b)] | 5 /<br>Syntax | Begriffs- u. Definitionsprofile (Online-Enzyklopädien, Wordnet, WWW) /<br>ROUGE für automatische Evaluation |
| *Information Desk*<br>[Xu u. a. (2005, 2006)] | 3 /<br>Wortf. | Begriffsprofile (k.A. zur Quelle) /<br>Wörter aus Begriffsprofil als Features für Ranking |
| *Fudan University*<br>[Zhang u. a. (2005)] | keine /<br>k.A. | Begriffsprofile (Online-Quellen; Fallback: alle Treffer) /<br>Extraktion mittels Begriffsprofilen |
| *UChongqing*<br>[Chen u. a. (2006)] | keine /<br>Stämme | Begriffsprofile (WWW); Schlüsselwörter zu Definitionstypen /<br>*Query expansion* in der IR-Phase; Extraktion nach Ranking mit Bigramm-Modell aus Begriffsprofil |

**CLEF**

| *System* | *Muster / Analysen* | *Zusatzinformation / Besonderheiten* |
|---|---|---|
| *Joost*<br>[Bouma u. a. (2006)] | 5 /<br>Syntax | *is-a*-Relationen aus Wikipedia-Definitionen /<br>Sprache: nl; Identifikation der Wikipediadefinitionen: Vgl. Fahmi und Bouma (2006) |
| *INAOE-UPV*<br>[Denicia-Carral u. a. (2006)] | k.A./<br>k.A. | Sprache: es |

**WWW**

| *System* | *Muster / Analysen* | *Zusatzinformation / Besonderheiten* |
|---|---|---|
| *DefScriber (s.o.)* | 18 / Syntax | Definitionsprofile (aus allen Treffern); vordefinierte Merkmale / Evaluation durch Fragebogen[2] |
| *WebQA / Mdef-WQA* [Figueroa und Neumann (2007); Figueroa (2008b,a)] | 8 / Wortformen | Sprachen: en, es; flexibles Matching (Überlappung) |
| *USingapore* [Cui u. a. (2005, 2004a)] | k.A. / Chunks | Begriffsprofile (aus Kookurrenzen in der Textgrundlage) / Suchmuster aus Beispielen generalisiert (durch Ersetzung aus Begriffsprofilen); flexibles Matching (Überlappung) |

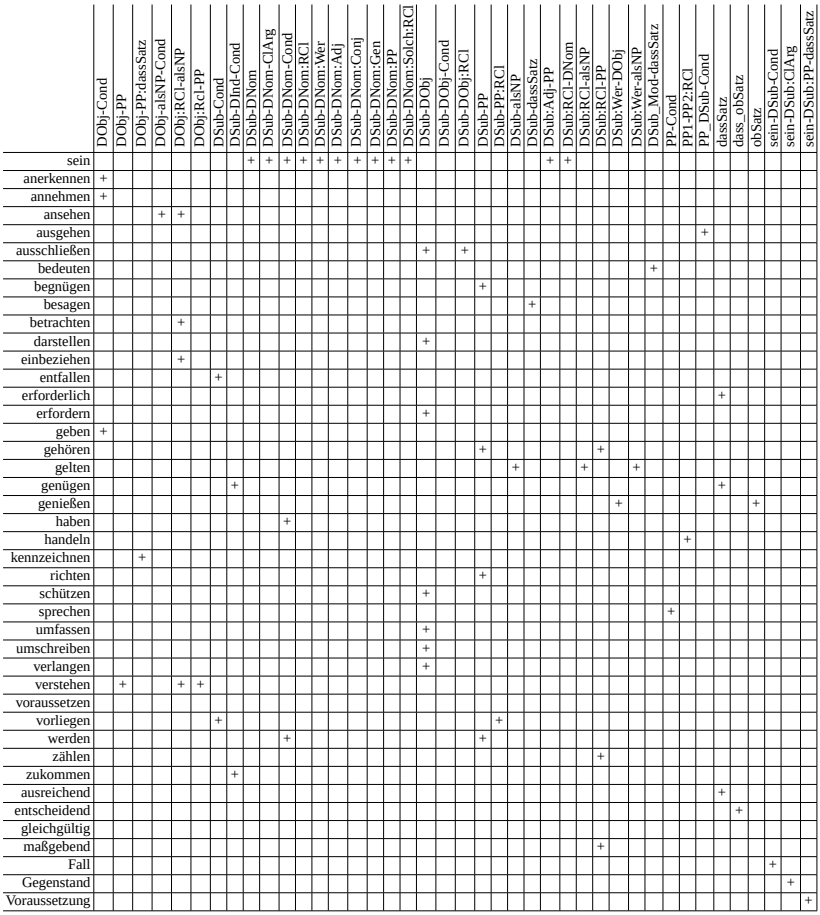
**Spezielle Systeme / einzelne Aspekte**

| *System* | *Muster / Analysen* | *Korpus / Besonderheiten* |
|---|---|---|
| *DEFQA* [Miliaraki und Androutsopoulos (2004)] | 8 / k.A. | Aquaint / vergleichen Ranking-Methoden anhand verschiedener Merkmale (z.B. Suchmuster, Hyperonyme, Position) |
| *MedQA* [Yu u. a. (2007)] | k.A. / Chunks | Medline, WWW / anwendungsnahes System; Evaluation: Usability (*think-aloud*-Protokolle und Fragebogen) |
| *Joost* [Fahmi und Bouma (2006)] | keine / k.A. | Wikipedia (Medizinische Artikel) / Extraktion durch Klassifikation anhand (hauptsächlich) syntaktischer Merkmale; Sprache: nl |

[1] Im Rahmen von TREC 12. Systeme ohne Rank-Angabe waren 2003 nicht am offiziellen Wettbewerb beteiligt.

[2] mit den Kriterien *Struktur, Verständlichkeit, Redundanz, Relevanz und Abdeckung*

# Anhang C: Dependenzbasierte Muster




**Erläuterungen**

| | | | | | |
|---|---|---|---|---|---|
| **:** | Einbettung | **PP** | Präpositionalphrase | **ClArg** | Klausales Argument (z.B. *dass*-Satz, erw. Inf.) |
| **_** | Disjunktion | **dassSatz** | *dass*-Satz | **Solch** | Attribut *solcher / solche / solches* |
| **DSub** | Tiefensubjekt | **obSatz** | *ob*-Satz | **RCl** | Relativsatz |
| **DObj** | Tiefenobjekt | **Conj** | Koordination | **Solch** | Attribut *ssolcher / solche / solches* |
| **DInd** | Indir. Objekt | **Gen** | Genitivattribut | **Cond** | Konditionalsatz |
| **DNom** | Prädikatsnomen | **Adj** | Adjektivattribut | **Wer** | *Was* bzw. *wer* zur Einleitung freier Relativsätze |

# Anhang D: Fehlerklassen (Analyse von Extraktionsfehlern)

| **A Recall-Fehler** | | |
|---|---|---|
| *Fehlerklasse* | *Beschreibung* | *Mögliche Ursachen / Beispiele* |
| A1 Satzgrenzenerkennung | | Satzzeichen in Abkürzungen und Normverweisen |
| A2 Keine Analyse | | Ressourcenprobleme, falsch kodierte Eingabe |
| A3 Keine PreDS | | Programm- / Ressourcenprobleme |
| A4 Parsefehler (Topo) | | Probleme der topologischen Grammatik: Komplexe Satzgefüge, ibs. Ambiguitäten zw. Subjunktion /Konjunktion und anderen Wortklassen |
| A5 Parsefehler (PreDS) | | Fehler der Konstruktionsheuristik, Probleme des Unterspezifikationsmechanismus |
| A6 Suchmuster zu eng | Satz wäre bei Weglassen eines klar eingrenzbaren Kriteriums in einem Suchmuster gefunden worden | Definition mit Kausalsatz, Suchmuster verlangt Konditionalsatz |
| A7 Fehlender *frame* | Satz gehört nicht zu den intendierten Treffern eines der vorhandenen Suchmuster (es gibt zwar Suchmuster mit dem entsprechenden Prädikat, aber keines mit einem passenden *frame*) | |
| A8 Fehlendes Prädikat | Satz gehört nicht zu den intendierten Treffern eines der vorhandenen Suchmuster (keine Suchmuster mit dem entsprechenden Prädikat) | |

| **B Präzisions-Fehler** | | |
|---|---|---|
| *Fehlerklasse* | *Beschreibung* | *Mögliche Ursachen / Beispiele* |
| B1 Fehlanalyse | Das Muster passt auf den Satz nur aufgrund eines Fehlers in der linguistischen Analyse | Konditionalsatz fälschlich einem Definitionsprädikat zugeordnet |
| B2 Suchmuster zu weit | Satz wäre aufgrund ergänzender verallgemeinerbarer Kriterien anhand der PreDS-Struktur als nicht-definitorisch klassifizierbar | |
| B3 Ausnahme | Im Satz ist eine im allgemeinen definitorische Formulierung aufgrund eines bestimmten Einzelaspekts nicht definitorisch | Nicht definierbares Konzept (“Stoppwort”) als Definiendum (z.B. Normverweis) |
| B4 Systematische Ambiguität | Bei semantisch ambigen Definitionsmustern: Muster tritt im Satz in nicht-definitorischer Bedeutung auf | *verlangen* mit belebtem Subjekt (und somit nicht zur Angabe eines Normgehalts o.ä.) |
| B5 Kontingente Aussage | Eindeutig nicht-definitorische Verwendung einer Formulierung ohne systematische Ambiguität | |
| B6 Subsumtion | Satz verwendet ein definitorisches Formulierungsmuster, um die Subsumtion eines konkreten Sachverhalts im konkreten Einzelfall zu Ausdruck zu bringen | Aussagen mit indexikalischen Elementen (“hier”, “dieser”) |
| B7 Zweifelsfall | Satz erfüllt semantische oder juristische Definitionskriterien nicht in ausreichendem Maße | zu geringer Abstraktionsgrad, zu stark auf spezielle Fälle eingeschränkt |

# Anhang E: Merkmale für Klassifikation und Ranking

| Merkmal | Beschreibung / Anmerkungen | Motivation |
|---|---|---|
| **Lexikalisch** | | |
| *Stoppwörter* | Manuell kompilierte und automatisch erweiterte Listen | Vgl. 6.1.1 |
| *Boostwort* | Treten im Satz Wörter auf, die aufgrund ihrer Bedeutung in Definitionen zu erwarten sind, aber nicht von unseren Suchmustern abgedeckt werden? | In vielen Definitionen wird (unabhängig vom Definitionsprädikat) explizit auf den Definitionscharakter der Aussage hingewiesen (z.B. durch Einbettung des Definiendum unter "transparente" Ausdrücke wie *Merkmal* oder *Begriff*) |
| *Subsumtionssignal* | Liste von Signalwörtern, die Subsumtion anzeigen. | Vgl. 6.1.1 |
| *Negation* | Treten im Satz Negationssignale auf (*nicht* auf Satzebene, *kein*, *un-* im Definiendum)? | Negationen sind in Definitionen zwar prinzipiell nicht ausgeschlossen. Vollwertige Definitionen werden jedoch meistens positiv formuliert, Negationen können den definitorischen Wert einer Aussage einschränken (vgl. Kap. 2) |
| *Modifikator* | Enthält die Definition modifizierende Satzadverbien? | Bestimmte Adverbien werden typischerweise in kontextualisierten Definitionen zur Einbettung in die Argumentation und zur genauen Regelung der Geltung verwendet (vgl. Kap. 2) |
| *Lexikalische Ähnlichkeit der Bestandteile* | Kosinus-Ähnlichkeit der Wortvektoren für (lemmatisiertes) Definiendum und Definiens | Definiendum und Definiens können ein hohes Maß an lexikalischer Überlappung aufweisen (etwa bei komplexen Definienda wie *Abfälle zur Entsorgung*, die durch Einschränkung von Bestandteilen der Definiendum-Phrase definiert werden). |
| *Morphologische Ähnlichkeit der Bestandteile* | Kosinus-Ähnlichkeit der Vektoren für Definiendum und Definiens unter Verwendung der identifizierten morphologischen Wurzeln anstelle der Lemmata (vgl. 4.2) | Durch die morphologische Zerlegung können z.B. Übergänge zwischen Wortklassen und Definitionen von Komposita auf der Basis eines Simplex berücksichtigt werden (z.B. *Abfallentsorgung* aus *Abfall* und *Entsorgung*) |

| Merkmal | Beschreibung / Anmerkungen | Motivation |
|---|---|---|
| **Referentiell** | | |
| *Definites Definiendum* | Im Unterschied zum in 6.1.1 beschriebenen Ansatz wird hier gezielt in der als Definiendum identifizierten Textspanne gesucht. | Vgl. 6.1.1 |
| *Definiter Genusbegriff* | Vgl. 2.3. Gesucht wird (wenn vorhanden) in der vom PreDS-Parser als Prädikatsnomen identifizierten Textspanne (ausschließlich untergeordneter Nebensätze) | Im Unterschied zum Definiendum deuten definiter Artikel bzw. Demonstrativpronomen im Genusbegriff darauf hin, dass dieser eingeschränkt wird, also möglicherweise Teil einer Definition ist. Die Einschränkung liefert die für die definite Verwendung notwendige Einzigkeits-Bedingung. |
| *Anaphorisches Definiendum* | Tritt im Definiendum ein Personalpronomen auf? | Ein anaphorisches Definiendum kann darauf hindeuten, dass der Satz Teil eines kohärenten “Definitionskomplexes” ist. |
| *Oberbegriff des Definiendum im Definiens* | Tritt im Definiens ein Oberbegriff (Germanet-Hyperonym) des Definiendum auf? | Deutet auf das Vorliegen einer Definition *per genus et differentiam* (oder einer verwandten Struktur) hin. |
| *Synonym des Definiendum im Definiens* | Tritt im Definiens ein (Germanet-) Synonym des Definiendum auf? | Deutet auf eine Substitutionsdefinition durch direkte Synonymangabe oder eine umschreibende Definition hin. |
| **Strukturell** | | |
| *Definiendum vor Definiens* | Geht das Definiendum in der Oberflächenreihenfolge dem Definiens voraus? | Aus Gründen der Informationsstruktur ist diese Reihenfolge in Definitionen wahrscheinlich. |
| *Definition kein Hauptsatz* | Ist das Definitionsprädikat einem anderen Prädikat untergeordnet? | Eingebettete Definitionsprädikate deuten u.U. auf bloß angeführte oder in Erwägung gezogene Äußerungen hin, deren definitorischer Wert somit zumindest eingeschränkt wird. |

| Merkmal | Beschreibung | Motivation |
|---|---|---|
| **Dokumentstruktur** | | |
| *Relative Dokumentposition* | Relative Position des Satzes im Dokument (Anteil der vorangegangen Sätze an der gesamten Länge der Urteilsbegründung). | Aufgrund ihrer Wichtigkeit für die Argumentation werden Definitionen in vielen Fällen relativ früh im Dokument gegeben. |
| *Treffer im Vorkontext* | Wurden in den drei dem extrahierten Satz vorangehenden Sätzen weitere Definitionskandidaten identifiziert? | Deutet darauf hin, dass der Satz Teil eines kohärenten "Definitionskomplexes" ist. |
| *Treffer im Nachkontext* | Wurden in den drei dem extrahierten Satz folgenden Sätzen weitere Definitionskandidaten identifiziert? | Deutet darauf hin, dass der Satz Teil eines kohärenten "Definitionskomplexes" ist. |
| **Sonstige** | | |
| *Satzlänge* | Länge des Satzes im Verhältnis zur maximalen Satzlänge im Dokument. | |
| *Rechtsprechungszitat* | Wurde im Satz durch die *Named Entity*-Erkennung eine Zitatangabe erkannt? | Definitionen werden oft durch Verweise auf frühere Rechtsprechung belegt oder zu dieser ins Verhältnis gesetzt. |
| *TF / IDF-Wert für das Definiendum* | TF / IDF-Wert für das Kopfwort des Definiendum (Term frequency / inverse document frequency. TF / IDF gewichtet ein Wort nach seiner Häufigkeit im betrachteten Kontext, berücksichtigt dabei aber auch seine generelle Häufigkeit. Vgl. Spärck Jones (1972) und Salton und Buckley (1988). In unserem Fall wurde der Wert auf der Grundlage der Auftreten in einem Vor- und Nachkontext von je drei Sätzen und der Dokumentfrequenz im CORTE-Korpus berechnet). | Definitionen beziehen sich mit großer Wahrscheinlichkeit auf im Kontext besonders prominente Begriffe. |
| *Präzisionsschätzung* | Vgl. 6.1. | |

# Literaturverzeichnis

[Abelson 1967] ABELSON, R.: Definition. In: EDWARDS, P. (Hrsg.): *Encyclopedia of Philosophy* Bd. 2. New York: MacMillan, 1967, S. 315–324

[Abney 1996] ABNEY, S.: Partial parsing via finite-state cascades. In: *Natural Language Engineering* 2 (1996), Nr. 4, S. 337–344

[Agichtein und Gravano 2000] AGICHTEIN, E.; GRAVANO, L.: Snowball: Extracting Relations from Large Plain-Text Collections. In: *Proceedings of the Fifth ACM International Conference on Digital Libraries*, 2000

[Ahlswede und Evens 1988] AHLSWEDE, T.; EVENS, M. W.: Parsing vs. Text Processing in the Analysis of Dictionary Definitions. In: *ACL*, 1988, S. 217–224

[Ahn u. a. 2004] AHN, D.; JIJKOUN, V.; MISHNE, G.; MÜLLER, K.; RIJKE, M. de; SCHLOBACH., S.: Using Wikipedia at the TREC QA Track. In: *Proceedings of TREC 2004*, 2004

[Alexy 1996] ALEXY, R.: *Theorie der juristischen Argumentation. Die Theorie des rationalen Diskurses als Theorie der juristischen Begründung*. 2. Frankfurt am Main: Suhrkamp, 1996

[Alshawi 1987] ALSHAWI, H.: Processing dictionary definitions with phrasal pattern hierarchies. In: *Computationlal Linguistics* 13 (1987), Nr. 3-4, S. 195–202

[Androutsopoulos und Galanis 2005] ANDROUTSOPOULOS, I.; GALANIS, D.: A Practically Unsupervised Learning Method to Identify Single-Snippet Answers to Definition Questions on the Web. In: (HLT/EMNLP, 2005)

[Anscombe 1958] ANSCOMBE, G. E. M.: On Brute Facts. In: *Analysis* 18 (1958), S. 69–72

[Appelt und Israel 1999] APPELT, D. E.; ISRAEL, D.: *Introduction to Information Extraction Technology*. Web document, `http://www.ai.sri.com/~appelt/ie-tutorial/`, 18. June 2010. 1999. – Tutorial held at the IJCAI 1999

[Arntz u. a. 2004] ARNTZ, R.; PICHT, H.; MAYER, F.: *Einführung in die Terminologiearbeit*. 5. Hildesheim ; Zürich: Olms, 2004

[Baeza-Yates und Ribeiro-Neto 1999] BAEZA-YATES, R. A.; RIBEIRO-NETO, B.: *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999

[Barnbrook 2002] BARNBROOK, G.: *Defining Language - A local grammar of definition sentences*. Amsterdam, 2002

[Barrière 2004a] BARRIÈRE, C.: Building a concept hierarchy from corpus analysis. In: *Terminology* 10 (2004), Nr. 2, S. 241–263

[Barrière 2004b] BARRIÈRE, C.: Knowledge-Rich Contexts Discovery. In: TAWFIK, A. Y. (Hrsg.); GOODWIN, S. D. (Hrsg.): *Proceedings of the Canadian Conference on AI 2004* Bd. 3060, Springer, 2004, S. 187–201

[Becker und Hayes 1963] BECKER, J.; HAYES, R. M.: *Information Storage and Retrieval: tools, elements, theories*. New York, 1963

[Benjamins u. a. 2005] BENJAMINS, V. R. (Hrsg.); CASANOVAS, P. (Hrsg.); BREUKER, J. (Hrsg.); GANGEMI, A. (Hrsg.): *Law and the Semantic Web: Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications [outcome of the Workshop on Legal Ontologies and Web-Based Legal Information Management, June 28, 2003, Edinburgh, UK & International Seminar on Law and the Semantic Web, November 20-21, 2003, Barcelona, Spain]*. Bd. *3369*. 2005

[Biagioli u. a. 2005] BIAGIOLI, C.; FRANCESCONI, E.; PASSERINI, A.; MONTEMAGNI, S.; SORIA, C.: Automatic semantics extraction in law documents. In: *Proceedings of the ICAIL 2005*. Bologna, Italy: ACM, 2005, S. 133–140

[Biasiotti u. a. 2008] BIASIOTTI, M.; FRANCESCONI, E.; PALMIRANI, M.; SARTOR, G.; VITALI, F.: Legal Informatics & Management of Legislative Documents / Global Centre for ICT in Parliament Working Paper No. 2. January 2008. – Forschungsbericht

[Bing 1987a] BING, J.: Designing text retrieval systems for conceptual searching. In: *Proceedings of the ICAIL 1987*. Boston, Massachusetts, United States: ACM, 1987, S. 43–51

[Bing 1987b] BING, J.: Performance of Legal Text Retrieval Systems: The Curse of Boole. In: *Law Librarian Journal* 79 (1987), S. 187–202

[Bix 1995] BIX, B.: *Law, Language, and Legal Determinacy*. Oxford, 1995

[Blair-Goldensohn u. a. 2003a] BLAIR-GOLDENSOHN, S.; MCKEOWN, K.; SCHLAIKJER, A. H.: DefScriber: a hybrid system for definitional QA. In: *Proceedings of the SIGIR 2003*, ACM, 2003, S. 462

[Blair-Goldensohn u. a. 2003b] BLAIR-GOLDENSOHN, S.; MCKEOWN, K.; SCHLAIKJER, A. H.: A Hybrid Approach for Answering Definitional Questions, CUCS-006-03. Columbia University, 2003. – Forschungsbericht

[Blair-Goldensohn u. a. 2003c] BLAIR-GOLDENSOHN, S.; MCKEOWN, K.; SCHLAIKJER, A. H.: A Hybrid Approach for QA Track Definitional Questions. In: *TREC*, 2003, S. 185–192

[Boer u. a. 2001] BOER, A.; WINKELS, R.; HOEKSTRA, R.: The CLIME Ontology. In: WINKELS, R. (Hrsg.): *Proceedings of the Second International Workshop on Legal Ontologies (LEGONT)*. Amsterdam, Netherlands, 2001, S. 37–47

[Boer u. a. 2004] BOER, A.; WINKELS, R.; ENGERS, T. M. van; MAAT, E. de: A Content Management System Based on an Event-based Model of Version Management Information in Legislation. In: (Gordon, 2004)

[Boer u. a. 2003] BOER, A.; WINKELS, R.; HOEKSTRA, R.; ENGERS, T. M. van: Knowledge Management for Legislative Drafting in an International Setting. In: BOURCIER, D. (Hrsg.): *Legal Knowledge and Information Systems. Jurix 2003: The Sixteenth Annual Conference.* Amsterdam: IOS Press, 2003 (Frontiers in Artificial Intelligence and Applications), S. 91–100

[Boguraev und Briscoe 1989] BOGURAEV, B.; BRISCOE, E.: Utilising the LDOCE grammar codes. (1989), S. 85–116

[Boguraev und Pustejovsky 1990] BOGURAEV, B.; PUSTEJOVSKY, J.: Lexical ambiguity and the role of knowledge representation in lexicon design. In: *Proceedings of COLIG 1993*. Helsinki, Finland: Association for Computational Linguistics, 1990, S. 36–41

[Bohrer 1993] BOHRER, A.: *Entwicklung eines internetgestützten Expertensystems zur Prüfung des Anwendungsbereichs urheberrechtlicher Abkommen*, Universität des Saarlandes, Dissertation, 1993

[Bouma u. a. 2006] BOUMA, G.; FAHMI, I.; MUR, J.; NOORD, G. van; PLAS, L. van der; TIEDEMANN, J.: Using Syntactic Knowledge for QA. In: PETERS, C. (Hrsg.); CLOUGH, P. (Hrsg.); GEY, F. C. (Hrsg.); KARLGREN, J.

(Hrsg.); MAGNINI, B. (Hrsg.); OARD, D. W. (Hrsg.); RIJKE, M. de (Hrsg.); STEMPFHUBER, M. (Hrsg.): *CLEF* Bd. 4730, Springer, 2006, S. 318–327

[Brants u. a. 2002] BRANTS, S.; DIPPER, S.; HANSEN, S.; LEZIUS, W.; SMITH, G.: The TIGER Treebank. In: *Proceedings of the Workshop on Treebanks and Linguistic Theories*. Sozopol, 2002

[Braun 1999] BRAUN, C.: *Flaches und robustes Parsen deutscher Satzgefüge*, Saarland University, Diplomarbeit, 1999

[Braun 2003] BRAUN, C.: Parsing German text for syntacto-semantic structures. In: *Prospects and Advances in the Syntax/Semantics Interface, Lorraine-Saarland Workshop Series*. Nancy, France, 2003, S. 99–102

[Breuker und Hoekstra 2004] BREUKER, J.; HOEKSTRA, R.: Epistemology and ontology in core ontologies: FOLaw and LRI-Core, two core ontologies for law. In: *Proceedings of EKAW Workshop on Core ontologies*, CEUR, 2004

[Brill u. a. 2001] BRILL, E.; LIN, J. J.; BANKO, M.; DUMAIS, S. T.; NG, A. Y.: Data-Intensive Question Answering. In: *TREC*, 2001

[Brin und Page 1998] BRIN, S.; PAGE, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: *Computer Networks* 30 (1998), Nr. 1-7, S. 107–117

[Briscoe 1989] BRISCOE, T.; BOGURAEV, B. (Hrsg.): *Computational lexicography for natural language processing*. White Plains, NY, USA: Longman Publishing Group, 1989

[Briscoe und Carroll 2000] BRISCOE, T.; CARROLL, J.: *Grammatical Relation annotation*. On-line document, `http://www.cogs.susx.ac.uk/lab/nlp/carroll/grdescription/index.html`, 16.6.2010. 2000

[Briscoe und Copestake 1999] BRISCOE, T.; COPESTAKE, A.: Lexical rules in constraint-based grammars. In: *Computationlal Linguistics* 25 (1999), Nr. 4, S. 487–526

[Briscoe u. a. 1993] BRISCOE, T. (Hrsg.); COPESTAKE, A. (Hrsg.); PAIVA, V. de (Hrsg.): *Inheritance, defaults and the lexicon*. New York, NY, USA: Cambridge University Press, 1993

[Briscoe u. a. 1990] BRISCOE, T.; COPESTAKE, A. A.; BOGURAEV, B.: Enjoy the Paper: Lexicology. In: *COLING*, 1990, S. 42–47

[Brüninghaus und Ashley 2005] BRÜNINGHAUS, S.; ASHLEY, K. D.: Reasoning with Textual Cases. In: MUÑOZ-AVILA, H. (Hrsg.); RICCI, F. (Hrsg.): *Proceedings of the ICCBR 2005* Bd. 3620, Springer, 2005, S. 137–151

[Büchel und Weber 1995] BÜCHEL, G.; WEBER, N.: Semantische Relationen in Definitionsstrukturen. In: HITZENBERGER, L. (Hrsg.): *Proceedings der GLDV-Jahrestagung 1995*, Georg Olms Verlag, 1995, S. 127–140

[Buitelaar u. a. 2005] BUITELAAR, P. (Hrsg.); CIMIANO, P. (Hrsg.); MAGNINI, B. (Hrsg.): *Frontiers in Artificial Intelligence and Applications Series*. Bd. 123: *Ontology Learning from Text: Methods, Evaluation and Applications*. Amsterdam: IOS Press, 7 2005

[Burhans 2002] BURHANS, D. T.: *A Question Answering Interpretation of Resolution Refutation*. Buffalo, State University of New York at Buffalo, Dissertation, 2002

[Busse 1992] BUSSE, D.: *Reihe Germanistische Linguistik*. Bd. 131: *Recht als Text. Linguistische Untersuchungen zur Arbeit mit Sprache in einer gesellschaftlichen Institution.* Tübingen: Max Niemeyer Verlag, 1992

[Calzolari 1984] CALZOLARI, N.: Detecting Patterns in a Lexical Data Base. In: *COLING*, 1984, S. 170–173

[Capper und Susskind 1988] CAPPER, P.; SUSSKIND, R.: *Latent Damage Law - The expert System*. Butterworths, 1988

[Carbonell und Goldstein 1998] CARBONELL, J.; GOLDSTEIN, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: *Proceedings of SIGIR 1998*. Melbourne, Australia: ACM, 1998, S. 335–336

[Carnap 1956] CARNAP, R. (Hrsg.): *Meaning and Necessity*. 2. Chicago, London, 1956

[Casanovas u. a. 2008] CASANOVAS, P. (Hrsg.); BIASIOTTI, M. A. (Hrsg.); FRANCESCONI, E. (Hrsg.); SAGRI, M.-T. (Hrsg.): *Proceedings of the 2nd Workshop on Legal Ontologies and Artificial Intelligence Techniques June 4th, 2007, Stanford University, Stanford, CA, USA*. Bd. *321*. CEUR-WS.org, 2008. (CEUR Workshop Proceedings)

[Chamberlin und Robie 2008] CHAMBERLIN, D. D.; ROBIE, J.: *XQuery 1.1*. World Wide Web Consortium, Working Draft WD-xquery-11-20081203. December 2008

[Charniak 1999] CHARNIAK, E.: A Maximum-Entropy-Inspired Parser, 1999, S. 132–139

[Chen u. a. 2006] CHEN, Y.; ZHOU, M.; WANG, S.: Reranking Answers for Definitional QA Using Language Modeling. In: *Proceedings of the ACL 2006*. Sidney, Australia: The Association for Computer Linguistics, 2006

[Chinchor 1998] CHINCHOR, N.: Overview of MUC-7. In: *Proceedings of the MUC-7, 1998*. Fairfax, VA, 1998

[Christ 1994] CHRIST, O.: A Modular and Flexible Architecture for an Integrated Corpus Query System. In: *Proceedings of COMPLEX 1994*. Budapest, Hungary, 1994

[Cimiano 2006] CIMIANO, P.: *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer, 2006

[Clark und DeRose 1999] CLARK, J.; DEROSE, S.: XML Path Language (XPath) version 1.0 / World Wide Web Consortium. November 1999. – Recommendation. See `http://www.w3.org/TR/xpath.html`

[Cleverdon 1972] CLEVERDON, C. W.: On the inverse relationship of recall and precision. In: *Journal of Documentation* 28(3) (1972), S. 195–201

[Cleverdon u. a. 1966] CLEVERDON, C. W.; MILLS, J.; KEEN, E. M.: *Factors determining the performance of indexing systems*. Cranfield, U.K: College of Aeronautics, 1966

[Cleverdon 1970] CLEVERDON, C.: Progress in documentation, evaluation tests of information retrieval systems. In: *Journal of Documentation* 26 (1970), S. 55–67

[Cohen 1960] COHEN, J.: A coefficient of agreement for nominal scales. In: *Educational and Psychological Measurement* 20 (1960), S. 37–46

[Crysmann u. a. 2002] CRYSMANN, B.; FRANK, A.; KIEFER, B.; MUELLER, S.; NEUMANN, G.; PISKORSKI, J.; SCHÄFER, U.; SIEGEL, M.; USZKOREIT, H.; XU, F.; BECKER, M.; KRIEGER, H.-U.: An Integrated Architecture for Shallow and Deep Processing. In: *ACL*, 2002, S. 441–448

[Cui u. a. 2004a] CUI, H.; KAN, M.-Y.; CHUA, T.-S.: Unsupervised learning of soft patterns for generating definitions from online news. In: FELDMAN, S. I. (Hrsg.); URETSKY, M. (Hrsg.); NAJORK, M. (Hrsg.); WILLS, C. E. (Hrsg.): *Proceedings of WWW 2004*, ACM, 2004, S. 90–99

[Cui u. a. 2005] CUI, H.; KAN, M.-Y.; CHUA, T.-S.: Generic soft pattern models for definitional question answering. In: BAEZA-YATES, R. A. (Hrsg.); ZIVIANI, N. (Hrsg.); MARCHIONINI, G. (Hrsg.); MOFFAT, A. (Hrsg.); TAIT, J. (Hrsg.): *Proceedings of the SIGIR 2005*, ACM, 2005, S. 384–391

[Cui u. a. 2004b] CUI, H.; LI, K.; SUN, R.; CHUA, T.-S.; KAN, M.-Y.: National University of Singapore at the TREC-13 Question Answering Main Task. In: *13th Text Retrieval Conference (TREC 2004)*, 2004

[Cunningham 2000] CUNNINGHAM, H.: *Software Architecture for Language Engineering*, University of Sheffield, Dissertation, 2000. – `http://gate.ac.uk/sale/thesis/`

[Daelemans u. a. 2009] DAELEMANS, W.; ZAVREL, J.; SLOOT, K. Van der; BOSCH, A. Van den: TiMBL: Tilburg Memory Based Learner, version 6.2, Reference Guide. / ILK Research Group. 2009 (09-01). – Forschungsbericht

[Daelemans und van den Bosch 2005] DAELEMANS, W.; BOSCH, A. van den: *Memory-Based Language Processing*. Cambridge University Press, 2005 (Studies in Natural Language Processing)

[Daum u. a. 2003] DAUM, M.; FOTH, K. A.; MENZEL, W.: Constraint Based Integration of Deep and Shallow Parsing Techniques. In: *Proceedings of the EACL 2003*, 2003, S. 99–106

[De Boni 2004] DE BONI, M.: *A relevance-based theoretical foundation for question answering*, University of York, Department of Computer Science, Dissertation, 2004

[De Mulder und van Noortwijk 1994] DE MULDER, R. V.; NOORTWIJK, C. van: A System for Ranking Documents according to their Relevance to a (Legal) Concept. In: *RIAO '94*, 1994, S. 733–750

[Del Gaudio und Branco 2007] DEL GAUDIO, R.; BRANCO, A.: Automatic Extraction of Definitions in Portuguese: A Rule-Based Approach. In: NEVES, J. (Hrsg.); SANTOS, M. F. (Hrsg.); MACHADO, J. (Hrsg.): *Proceedings of the EPIA Workshop 2007* Bd. 4874, Springer, 2007, S. 659–670

[Denicia-Carral u. a. 2006] DENICIA-CARRAL, C.; GÓMEZ, M. M. y; PINEDA, L. V.; HERNÁNDEZ, R.: A Text Mining Approach for Definition Question Answering. In: SALAKOSKI, T. (Hrsg.); GINTER, F. (Hrsg.); PYYSALO, S. (Hrsg.); PAHIKKALA, T. (Hrsg.): *Proceedings of FinTAL 2006* Bd. 4139, Springer, 2006, S. 76–86

[Dietrich 2000] DIETRICH, R. u. K. K.: Transparent oder verständlich oder wie was verstanden wird - Eine empirische Untersuchung zum Verstehen eines juristischen Textes. In: *Zeitschrift für Literaturwissenschaft und Linguistik* (2000), Nr. 118, S. 67–95

[Dietrich und Schmidt 2002] DIETRICH, R.; SCHMIDT, C.: Zur Lesbarkeit von Verbrauchertexten. Ein Beitrag aus der Sicht der Textproduktion. In: *Sprache des Rechts II. Themenheft der Zeitschrift für Literaturwissenschaft und Linguistik* 32 (2002), Nr. 128, S. 34–62

[Dini u. a. 2005] DINI, L.; PETERS, W.; LIEBWALD, D.; SCHWEIGHOFER, E.; MOMMERS, L.; VOERMANS, W.: Cross-lingual legal information retrieval using a WordNet architecture. In: *Proceedings of the ICAIL 2005*. Bologna, Italy: ACM Press, 2005, S. 163–167

[Dipper 2003] DIPPER, S.: *Implementing and Documenting Large-Scale Grammars - German LFG*, Dissertation, 2003

[Drach 1937] DRACH, E.: *Grundgedanken der deutschen Satzlehre*. Frankfurt a. M.: Diesterweg, 1937

[Dubey 2005] DUBEY, A.: What to do when lexicalization fails: parsing German with suffix analysis and smoothing. In: *Proceedings of ACL 2005*. Ann Arbor, Michigan, 2005

[Echihabi u. a. 2003] ECHIHABI, A.; HERMJAKOB, U.; HOVY, E. H.; MARCU, D.; MELZ, E.; RAVICHANDRAN, D.: Multiple-Engine Question Answering in TextMap. In: *TREC*, 2003, S. 772–781

[Eichhoff-Cyrus und Antos 2008] EICHHOFF-CYRUS, K. M. (Hrsg.); ANTOS, G. (Hrsg.): *Thema Deutsch*. Bd. 9: *Verständlichkeit als Bürgerrecht? Die Rechts- und Verwaltungssprache in der öffentlichen Diskussion*. Mannheim/Leipzig/Wien/Zürich: Dudenverlag, 2008

[van Engers u. a. 2004] ENGERS, T. M. van; GOG, R. van; SAYAH, K.: A Case Study on Automated Norm Extraction. In: (Gordon, 2004)

[Engisch 1997] ENGISCH, K.: *Einführung in das juristische Denken*. Stuttgart, 1997

[Fahmi und Bouma 2006] FAHMI, I.; BOUMA, G.: Learning to Identify Definitions using Syntactic Features. In: *Proceedings of the EACL 2006 workshop on Learning Structured Information in Natural Language Applications.*, 2006

[Farzindar und Lapalme 2004a] FARZINDAR, A.; LAPALME, G.: Legal texts summarization by exploration of the thematic structures and argumentative roles. In: *Text Summarization Branches Out Conference held in conjunction with ACL 04*. Barcelona, Spain, jul 2004

[Farzindar und Lapalme 2004b] FARZINDAR, A.; LAPALME, G.: LetSum, an Automatic Legal Text Summarizing System. In: (Gordon, 2004)

[Feldman und Sanger 2006] FELDMAN, R.; SANGER, J.: *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, December 2006

[Fellbaum 1998] FELLBAUM, C. (Hrsg.): *Wordnet, an Electronic Lexical Database*. MIT Press, 1998

[Fickentscher 1977] FICKENTSCHER, W.: *Methoden des Rechts in vergleichender Darstellung*. Bd. 4. Tübingen, 1977

[Figueroa 2008a] FIGUEROA, A.: Boosting the Recall of descriptive Phrases in Web Snippets. In: *Proceedings of LangTech 2008*. Rome, Italy, 2008

[Figueroa 2008b] FIGUEROA, A.: Retrieving Answers to Definition Questions. In: *Proceedings of QAWEB 2008*, 2008

[Figueroa und Neumann 2007] FIGUEROA, A.; NEUMANN, G.: A Multilingual Framework for Searching Definitions on Web Snippets. In: HERTZBERG, J. (Hrsg.); BEETZ, M. (Hrsg.); ENGLERT, R. (Hrsg.): *KI* Bd. 4667, Springer, 2007, S. 144–159

[Fliedner 2001] FLIEDNER, G.: *Überprüfung und Korrektur von Nominalkongruenz im Deutschen*, Universität des Saarlandes, Diplomarbeit, 2001

[Fliedner 2002] FLIEDNER, G.: A system for checking NP agreement in German texts. In: *Proceedings of the ACL 2002, Student Workshop*, 2002

[Fliedner 2007] FLIEDNER, G.: *Saarbrücken Dissertations in Computational Linguistic and Language Technology*. Bd. XXIII: *Linguistically Informed Question Answering*. Saarbrücken: Universität des Saarlandes und DFKI GmbH, 2007

[Fliedner und Braun 2005] FLIEDNER, G.; BRAUN, C.: Richtlinien für die Annotation mittels grammatischer Relationen / Universität des Saarlandes. 2005. – Forschungsbericht

[Flowerdew 1992] FLOWERDEW, J.: Definitions in Science Lectures. In: *Applied Linguistics* 13(2) (1992), S. 202–221

[Freund 1999] FREUND, Y.: The alternating decision tree learning algorithm. In: *Machine Learning: Proceedings of the Sixteenth International Conference*, Morgan Kaufmann, 1999, S. 124–133

[Friedland u. a. 2004] FRIEDLAND, N. S.; ALLEN, P. G.; MATTHEWS, G.; WITBROCK, M. J.; BAXTER, D.; CURTIS, J.; SHEPARD, B.; MIRAGLIA, P.; ANGELE, J.; STAAB, S.; MÖNCH, E.; OPPERMANN, H.; WENKE, D.; ISRAEL, D. J.; CHAUDHRI, V. K.; PORTER, B. W.; BARKER, K.; FAN, J.; CHAW, S. Y.; YEH, P. Z.; TECUCI, D.; CLARK, P.: Project Halo: Towards a Digital Aristotle. In: *AI Magazine* 25 (2004), Nr. 4, S. 29–48

[Friedman u. a. 1998] FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R.: Additive Logistic Regression: a Statistical View of Boosting. In: *Annals of Statistics* 28 (1998), S. 2000

[Fujii und Ishikawa 2004] FUJII, A.; ISHIKAWA, T.: Summarizing Encyclopedic Term Descriptions on the Web. In: *CoRR* cs.CL/0407026 (2004)

[Gaizauskas u. a. 2003] GAIZAUSKAS, R. J.; GREENWOOD, M. A.; HEPPLE, M.; ROBERTS, I.; SAGGION, H.; SARGAISON, M.: The University of Sheffield's TREC 2003 Q&A Experiments. In: *TREC*, 2003, S. 782–790

[Gantner und Ebenhoch 2001] GANTNER, F.; EBENHOCH, P.: Der Saarbrücker Standard für Gerichtsentscheidungen (kommentierte Fassung). In: *JurPC Web-Dok.* 116 (2001)

[Giles u. a. 1998] GILES, C. L.; BOLLACKER, K. D.; LAWRENCE, S.: CiteSeer: an automatic citation indexing system. In: *Proceedings of the third ACM conference on Digital libraries 1998*. Pittsburgh, Pennsylvania, United States: ACM, 1998, S. 89–98

[Gordon 2004] GORDON, T. (Hrsg.): *7th Annual Conference on Legal Knowledge and Information Systems (JURIX 2004)*. IOS Press, 2004

[Gospodneti und Hatcher 2004] GOSPODNETI, O.; HATCHER, E.: *Lucene in Action*. Manning Publications, 2004

[Graff 2002] GRAFF, D.: *The AQUAINT Corpus of English News Text*. Philadelphia: , 2002

[Greve und Wentura 1997] GREVE, W. (Hrsg.); WENTURA, D. (Hrsg.): *Wissenschaftliche Beobachtung: Eine Einführung.* Weinheim, 1997

[Grover u. a. 2003] GROVER, C.; HACHEY, B.; HUGHSON, I.; KORYCINSKI, C.: Automatic Summarisation of Legal Documents. In: *Proceedings of ICAIL 2003*, 2003, S. 243–251

[Hachey und Grover 2004] HACHEY, B.; GROVER, C.: Sentence Classification Experiments for Legal Text Summarization. In: (Gordon, 2004)

[Hachey und Grover 2005] HACHEY, B.; GROVER, C.: Automatic legal text summarisation: experiments with summary structuring. In: *Proceedings of the ICAIL 2005*. Bologna, Italy: ACM, 2005, S. 75–84

[Hafner 1978] HAFNER, C. D.: *An information retrieval system based on a computer model of legal knowledge.* Ann Arbor, MI, USA, Dissertation, 1978

[Haft und Lehmann 1989] HAFT, F. (Hrsg.); LEHMANN, H. (Hrsg.): *Neue Methoden im Recht*. Bd. 6: *Das LEX-Projekt*. Tübingen: Attempto-Verlag, 1989

[Hall u. a. 2009] HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H.: The WEKA Data Mining Software: An Update. In: *SIGKDD Explorations* 11 (2009), Nr. 1

[Hamp und Feldweg 1997] HAMP, B.; FELDWEG, H.: GermaNet - a Lexical-Semantic Net for German. In: *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 1997, S. 9–15

[Han u. a. 2005] HAN, K.-S.; SONG, Y.-I.; KIM, S.-B.; RIM, H.-C.: Phrase-Based Definitional Question Answering Using Definition Terminology. In: LEE, G. G. (Hrsg.); YAMADA, A. (Hrsg.); MENG, H. (Hrsg.); MYAENG, S.-H. (Hrsg.): *Proceedings of AIRS 2005* Bd. 3689, Springer, 2005, S. 246–259

[Han u. a. 2006a] HAN, K.-S.; SONG, Y.-I.; KIM, S.-B.; RIM, H.-C.: A Definitional Question Answering System Based on Phrase Extraction Using Syntactic Patterns. In: *IEICE Transactions* 89-D (2006), Nr. 4, S. 1601–1605

[Han u. a. 2006b] HAN, K.-S.; SONG, Y.-I.; RIM, H.-C.: Probabilistic model for definitional question answering. In: *Proceedings of SIGIR 2006*. Seattle, Washington, USA: ACM, 2006, S. 212–219

[Hanks 1987] HANKS, P.: Definitions and Explanations. In: SINCLAIR, J. M. (Hrsg.): *Looking Up: Account of the Cobuild Project in Lexical Computing*. London, 1987, S. 117–136

[Hansen u. a. 2006] HANSEN, S.; DIRKSEN, R.; KÜCHLER, M.; KUNZ, K.; NEUMANN, S.: Comprehensible legal texts - utopia or a question of wording? On processing rephrased German court decisions. In: *Hermes - Journal of Language and Communication Studies* (2006), Nr. 36, S. 15–40

[Harabagiu u. a. 2005] HARABAGIU, S.; MOLDOVAN, D.; CLARK, C.; BOWDEN, M.; HICKL, A.; WANG, P.: Employing Two Question Answering Systems in TREC 2005. In: *Fourteenth Text REtrieval Conference*, 2005

[Harabagiu u. a. 2003] HARABAGIU, S. M.; MOLDOVAN, D. I.; CLARK, C.; BOWDEN, M.; WILLIAMS, J.; BENSLEY, J.: Answer Mining by Combining Extraction Techniques with Abductive Reasoning. In: *TREC*, 2003, S. 375–382

[Hare 1952] HARE, R. M.: *The Language of Morals*. London, Oxford, New York, 1952

[Hart 1961] HART, H. L. A.: *The Concept of Law*. Oxford: Clarendon Press, 1961

[Hearst 1992] HEARST, M. A.: Automatic acquisition of hyponyms from large text corpora. In: *Proceedings of COLING 1992*. Nantes, France: Association for Computational Linguistics, 1992, S. 539–545

[Hempel 1950] HEMPEL, C. G.: Problems and Changes in the Empiricist Criterion of Meaning. In: *Revue Internationale de Philosophie* 11 (1950), S. 41–63

[Hildebrandt u. a. 2004] HILDEBRANDT, W.; KATZ, B.; LIN, J. J.: Answering Definition Questions with Multiple Knowledge Sources. In: *Proceedings of HLT-NAACL 2004*, 2004, S. 49–56

[Hirschman und Gaizauskas 2001] HIRSCHMAN, L.; GAIZAUSKAS, R.: Natural language question answering: the view from here. In: *Natural Language Engineering* 7 (2001), Nr. 4, S. 275–300

[HLT/EMNLP 2005] HLT/EMNLP (Hrsg.): *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*. The Association for Computational Linguistics, 2005

[Hoekstra u. a. 2007] HOEKSTRA, R.; BREUKER, J.; BELLO, M. D.; BOER, A.: The LKIF Core Ontology of Basic Legal Concepts. In: (Casanovas u. a., 2008), S. 43–63

[Hovy u. a. 2003] HOVY, E. H.; PHILPOT, A.; KLAVANS, J.; GERMANN, U.; DAVIS, P. T.: Extending Metadata Definitions by Automatically Extracting and Organizing Glossary Definitions. In: *Proceedings of the National Conference on Digital Government Research*. Boston, Masschusetts, 2003

[Ilson 1986] ILSON, R.: General English Dictionaries for Foreign Learners: Explanatory Techniques in Dictionaries. In: *Lexicographica* 2 (1986), S. 215–221

[Jandach 1993] JANDACH, T.: *Juristische Expertensysteme. Methodische Grundlagen ihrer Entwicklung*. Berlin, 1993

[JCDL 2001] JCDL (Hrsg.): *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2001, Roanoke, Virginia, USA, June 24-28, 2001, Proceedings*. ACM, 2001

[Jijkoun u. a. 2003] JIJKOUN, V.; MISHNE, G.; MONZ, C.; RIJKE, M. de; SCHLOBACH, S.; TSUR, O.: The University of Amsterdam at the TREC 2003 Question Answering Track. In: *Proceedings of TREC 2003*, 2003, S. 586–593

[Jones 2007] JONES, K. S.: Automatic summarising: The state of the art. In: *Information Processing & Management* 43 (2007), Nr. 6, S. 1449–1481

[Jørgensen 1937] JØRGENSEN, J.: Imperatives and Logic. In: *Erkenntnis* (1937), S. 288 ff.

[Kao und Poteet 2006] KAO, A.; POTEET, S.: *Natural Language Processing and Text Mining*. Dordrecht: Springer, 2006

[Katz und Lin 2003] KATZ, B.; LIN, J.: Selectively using relations to improve precision in question answering. In: *Proceedings of the EACL-2003 Workshop on Natural Language Processing for Question Answering*, 2003

[Katz u. a. 2003] KATZ, B.; LIN, J. J.; LORETO, D.; HILDEBRANDT, W.; BILOTTI, M. W.; FELSHIN, S.; FERNANDES, A.; MARTON, G.; MORA, F.: Integrating Web-based and Corpus-based Techniques for Question Answering. In: *Proceedings of TREC 2003*, 2003, S. 426–435

[Kent 1971] KENT, A.: *Information Analysis and Retrieval*. New York, NY, USA: John Wiley & Sons, Inc., 1971

[Kilgarriff u. a. 2004] KILGARRIFF, A.; PAVEL RYCHLY, P.; SMRZ, P.; TUGWELL, D.: The sketch engine. In: *Proceedings of the EURALEX International Congress 2004*, 2004, S. 105–116

[Kingston u. a. 2004] KINGSTON, J.; SCHAFER, B.; VANDENBERGHE, W.: Towards a Financial Fraud Ontology: A Legal Modelling Approach. In: *Artificial Intelligence and Law* 12 (2004), Nr. 4, S. 419–446

[Klavans und Muresan 2001a] KLAVANS, J. L.; MURESAN, S.: Evaluation of DEFINDER: a system to mine definitions from consumer-oriented medical text. In: (JCDL, 2001)

[Klavans und Muresan 2001b] KLAVANS, J. L.; MURESAN, S.: Evaluation of the DEFINDER System for Fully Automatic Glossary Construction. In: *Proceedings of the American Medical Informatics Association Symposium 2001*, 2001

[Klavans und Muresan 2002] KLAVANS, J. L.; MURESAN, S.: A method for automatically building and evaluating dictionary resources. In: *Proceedings of the LREC 2002*, 2002

[Klavans und Whitman 2001] KLAVANS, J.; WHITMAN, B.: Extracting taxonomic relationships from on-line definitional sources using LEXING. In: (JCDL, 2001), S. 257–258

[Klein 2004] KLEIN, W.: Ein Gemeinwesen, in dem das Volk herrscht, darf nicht von Gesetzen beherrscht werden, die das Volk nicht versteht. In: LERCH, K. D. (Hrsg.): *Recht verstehen. Verständlichkeit, Missverständlichkeit und Unverständlichkeit von Recht*. Berlin, New York: Walter de Gruyter, 2004, S. 197–203

[Koch 2003] KOCH, H.-J.: Deduktive Entscheidungsbegründung. In: ALEXY, R. (Hrsg.); KOCH, H.-J. (Hrsg.); KUHLEN, L. (Hrsg.); RÜSSMANN, H. (Hrsg.): *Elemente einer juristischen Begründungslehre*. München: Nomos Verlagsgesellschaft, 2003, S. 37–60

[Koch und Rüßmann 1982] KOCH, H.-J.; RÜSSMANN, H.: *Juristische Begründungslehre. Eine Einführung in die Grundprobleme der Rechtswissenschaft*. München, 1982

[Kohavi 1996] KOHAVI, R.: Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid. In: *Proceedings of KDD 1996*, 1996, S. 202–207

[van Kralingen 1995] KRALINGEN, R. van: *Frame-based Conceptual Models of Statute Law*. Kluwer Law International, The Hague, The Netherlands, 1995 (Computer / Law Series)

[Krieger und Nerbonne 1993] KRIEGER, H.-U.; NERBONNE, J.: Feature-based inheritance networks for computational lexicons. (1993), S. 90–136

[Kriele 1967] KRIELE, M.: *Theorie der Rechtsgewinnung*. Berlin, 1967

[Lakoff 1973] LAKOFF, G.: Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts. In: *Journal of Philosophical Logic* 2 (1973), Nr. 4

[Lame 2002] LAME, G.: *Ontology based information retrieval in french legal documents*, Ecole nationale supérieure des mines de Paris, Centre de recherche en informatique, Dissertation, 2002

[Lame 2003] LAME, G.: Using NLP Techniques to Identify Legal Ontology Components: Concepts and Relations. In: (Benjamins u. a., 2005), S. 169–184

[Lame und Desprès 2005] LAME, G.; DESPRÈS, S.: Updating ontologies in the legal domain. In: *Proceedings of the ICAIL 2005*. Bologna, Italy: ACM, 2005, S. 155–162

[Landau 2001] LANDAU, S. I.: *Dictionaries – the art and craft of lexicography*. 2. Cambridge: Cambridge University Press, 2001

[Landis und Koch 1977] LANDIS, J.; KOCH, G.: The measurement of observer agreement for categorical data. In: *Biometrics* 33 (1977), S. 159–174

[Larenz 1991] LARENZ, K.: *Methodenlehre der Rechtswissenschaft*. Berlin, 1991

[Leary u. a. 2003] LEARY, R.; VANDENBERGHE, W.; ZELEZNIKOW, J.: Towards a Financial Fraud Ontology; A Legal Modelling Approach. In: *Proceedings of the Workshop on Legal Ontologies & Web Based Legal Information Managment at ICAIL 2003*, 2003

[Lenci u. a. 2007] LENCI, A.; MONTEMAGNI, S.; PIRRELLI, V.; VENTURI, G.: NLP-based Ontology Learning from Legal Texts. A Case Study. In: (Casanovas u. a., 2008), S. 113–129

[Lezius 2002] LEZIUS, W.: *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora*, IMS, University of Stuttgart, Dissertation, December 2002. – Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), volume 8, number 4

[Liebwald 2007] LIEBWALD, D.: Semantic Spaces and Multilingualism in the Law: The Challenge of Legal Knowledge Management. In: (Casanovas u. a., 2008), S. 131–148

[Lin und Hovy 2003] LIN, C.-Y.; HOVY, E. H.: Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In: *Proceedings of - NAACL 2003*, 2003

[Lin und Pantel 2002] LIN, D.; PANTEL, P.: Concept Discovery from Text. In: *Proceedings of COLING 2002*, 2002

[Lin und Demner-Fushman 2005] LIN, J. J.; DEMNER-FUSHMAN, D.: Automatically Evaluating Answers to Definition Questions. In: (HLT/EMNLP, 2005)

[Liu u. a. 2003] LIU, B.; CHIN, C. W.; NG, H. T.: Mining topic-specific concepts and definitions on the web. In: *Proceedings of WWW 2003*, 2003, S. 251–260

[Loevinger 1949] LOEVINGER, L.: Jurimetrics: The Next Step Forward. In: *Minnesota Law Review* 33 (1949), April, Nr. 5, S. 455–93

[MacCormick und Weinberger 1986] MACCORMICK, N.; WEINBERGER, O.: *An Institutional Theory of Law*. Dordrecht: Reidel, 1986

[MacFarquhar und Richards 1983] MACFARQUHAR, P. D.; RICHARDS, J. C.: On Dictionaries and Definitions. In: *RELC Journal* 14 (1983), Nr. 1, S. 111–124

[Magnini u. a. 2004] MAGNINI, B.; VALLIN, A.; AYACHE, C.; ERBACH, G.; PEÑAS, A.; RIJKE, M. de; ROCHA, P.; SIMOV, K. I.; SUTCLIFFE, R. F. E.: Overview of the CLEF 2004 Multilingual Question Answering Track. In: PETERS, C. (Hrsg.); CLOUGH, P. (Hrsg.); GONZALO, J. (Hrsg.); JONES, G. J. F. (Hrsg.); KLUCK, M. (Hrsg.); MAGNINI, B. (Hrsg.): *CLEF* Bd. 3491, Springer, 2004, S. 371–391

[Malaise u. a. 2005] MALAISE, V.; ZWEIGENBAUM, P.; BACHIMONT, B.: Mining defining contexts to help structuring differential ontologies. In: *Terminology* 11 (2005), Nr. 1, S. 21–53

[Malaisé u. a. 2004] MALAISÉ, V.; ZWEIGENBAUM, P.; BACHIMONT, B.: Detecting Semantic Relations between Terms in Definitions. In: ANANADIOU, S. (Hrsg.); ZWEIGENBAUM, P. (Hrsg.): *Proceedings of CompuTerm 2004: 3rd International Workshop on Computational Terminology at COLING 2004*. Geneva, Switzerland: COLING, August 29 2004, S. 55–62

[Mani 2001] MANI, I.: *Automatic Summarization*. Amsterdam/Philadephia: John Benjamins Publishing Company, 2001

[Manning und Schütze 1999] MANNING, C. D.; SCHÜTZE, H.: *Foundations of statistical natural language processing*. Cambridge, MA, USA: MIT Press, 1999

[Markowitz u. a. 1986] MARKOWITZ, J.; AHLSWEDE, T.; EVENS, M.: Semantically significant patterns in dictionary definitions. In: *Proceedings of ACL 1986*. Morristown, NJ, USA: Association for Computational Linguistics, 1986, S. 112–119

[Maxwell und Kaplan 1991] MAXWELL, J. T.; KAPLAN, R. M.: A Method for Disjunctive Constraint Satisfaction. In: *Current Issues in Parsing Technology*. 1991

[McCarty 1977] MCCARTY, L. T.: Reflections on TAXMAN: An Experiment in Artificial Intelligence and Legal Reasoning. In: *Harvard Law Review* 90 (1977), March, Nr. 5, S. 837–93

[McCarty 1980] MCCARTY, L. T.: The TAXMAN Project: Towards a Cognitive Theory of Legal Argument. In: NIBLETT, B. (Hrsg.): *Computer Science and Law*. Cambridge: Cambridge University Press, 1980, Kap. 3, S. 23–43

[McCarty 1989] MCCARTY, L. T.: A Language for Legal Discourse I. Basic Fatures. In: *Proceedings of the ICAIL 1989*, 1989

[McCarty 2007] MCCARTY, L. T.: Deep semantic interpretations of legal texts. In: *Proceedings of the ICAIL 2007*. Stanford, California: ACM, 2007, S. 217–224

[Meyer 2001] MEYER, I.: Extracting knowledge-rich contexts for terminography. In: BOURIGAULT, D. (Hrsg.); JACQUEMIN, C. (Hrsg.); L'HOMME, M.-C. (Hrsg.): *Recent Advances in Computational Terminology*. Amsterdam, 2001, S. 279–302

[Miliaraki und Androutsopoulos 2004] MILIARAKI, S.; ANDROUTSOPOULOS, I.: Learning to Identify Single-Snippet Answers to Definition Questions. In: *Proceedings of COLING 2004*, 2004, S. 1360–1366

[Moens 2006] MOENS, M.-F.: *Information Extraction: Algorithms and Prospects in a Retrieval Context (The Information Retrieval Series)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006

[Moens u. a. 2007] MOENS, M.-F.; BOIY, E.; PALAU, R. M.; REED, C.: Automatic detection of arguments in legal texts. In: *Proceedings of the ICAIL 2007*. Stanford, California: ACM, 2007, S. 225–230

[Moens u. a. 1997] MOENS, M.-F.; UYTTENDAELE, C.; DUMORTIER, J.: Abstracting of legal cases: the SALOMON experience. In: *Proceedings of the ICAIL 1997*. Melbourne, Australia: ACM, 1997, S. 114–122

[Moldovan u. a. 2002] MOLDOVAN, D. I.; HARABAGIU, S. M.; GIRJU, R.; MORARESCU, P.; LACATUSU, V. F.; NOVISCHI, A.; BADULESCU, A.; BOLOHAN, O.: LCC Tools for Question Answering. In: *Proceedings of TREC 2002*, 2002

[Müller und Kasper 2000] MÜLLER, S.; KASPER, W.: HPSG Analysis of German. In: WAHLSTER, W. (Hrsg.): *Verbmobil. Foundations of Speech-to-Speech Translation*. Artificial Intelligence. Berlin, Germany: Springer, 2000, S. 238 – 253

[Muresan u. a. 2003] MURESAN, S.; POPPER, S. D.; DAVIS, P. T.; KLAVANS, J. L.: Building a Terminological Database from Heterogeneous Definitional Sources. In: *Proceedings of the National Conference on Digital Government Research 2003*. Boston, Masschusetts, 2003

[Müller 1994] MÜLLER, F.: *Strukturierende Rechtslehre*. Berlin, 1994

[Müller 1997] MÜLLER, F.: *Juristische Methodik*. Berlin, 1997

[Neff und Boguraev 1989] NEFF, M. S.; BOGURAEV, B.: Dictionaries, Dictionary Grammars and Dictionary Entry Parsing. In: *Proceedings of the ACL 1989*, 1989, S. 91–101

[Neumann 2009] NEUMANN, S.: Improving the comprehensibility of German court decisions. In: GREWENDORF, G. (Hrsg.); RATHERT, M. (Hrsg.): *Formal Linguistics and Law*. Berlin: Mouton de Gruyter, 2009 (Trends in Linguistics. Studies and Monographs (TiLSM))

[Neumann 2004] NEUMANN, U.: Juristische Logik. In: KAUFMANN, A. (Hrsg.); HASSEMER, W. (Hrsg.); NEUMANN, U. (Hrsg.): *Einführung in die Rechtsphilosophie und Rechtstheorie der Gegenwart*. Heidelberg, 2004, S. 298–318

[NIST 2006] NIST: Common Evaluation Measures. In: *Proceedings of TREC 2006*, NIST, 2006 (NIST Special Publication SP 500-272)

[Pado 2007] PADO, S.: *Cross-Lingual Annotation Projection Models for Role-Semantic Information*, Saarland University, Dissertation, 2007

[Pado und Lapata 2007] PADO, S.; LAPATA, M.: Dependency-based construction of semantic space models. In: *Computational Linguistics* 33 (2007), Nr. 2, S. 161–199

[Park u. a. 2002] PARK, Y.; BYRD, R. J.; BOGURAEV, B.: Automatic Glossary Extraction: Beyond Terminology Identification. In: *Proceedings of COLING 2002*, 2002

[Pearson 1998] PEARSON, J.: *Terms in context*. Amsterdam: Benjamins, 1998

[Peng u. a. 2005] PENG, F.; WEISCHEDEL, R. M.; LICUANAN, A.; XU, J.: Combining Deep Linguistics Analysis and Surface Pattern Learning: A Hybrid Approach to Chinese Definitional Question Answering. In: (HLT/EMNLP, 2005)

[Pinkal 1980] PINKAL, M.: Semantische Vagheit: Phänomene und Theorien, Teil 1. In: *Linguistische Berichte* 70 (1980), S. 1–26

[Platt 1998] PLATT, J.: Machines using Sequential Minimal Optimization. In: BURGES, C. (Hrsg.); SCHOELKOPF, A. S. (Hrsg.): *Advances in Kernel Methods - Support Vector Learning*. 1998

[Prager u. a. 2001] PRAGER, J.; RADEV, D.; CZUBA, K.: Answering What-Is Questions by Virtual Annotation. In: *Proceedings of HLT 2001*. San Diego, CA, 2001, S. 26–30

[Prager u. a. 2003] PRAGER, J. M.; CHU-CARROLL, J.; CZUBA, K.; WELTY, C. A.; ITTYCHERIAH, A.; MAHINDRU, R.: IBM's PIQUANT in TREC2003. In: *Proceedings of TREC 2003*, 2003, S. 283–292

[Przepiorkowski u. a. 2007] PRZEPIORKOWSKI, A.; DEGORSKI, L.; WOJTOWICZ, B.: On the evaluation of Polish definition extraction grammars. In: *Proceedings of the LTC 2007*. Poznan, Poland, 2007

[Pustejovsky 1995] PUSTEJOVSKY, J.: *The generative lexicon*. MIT Press, 1995

[Quine 1951] QUINE, W. V. O.: Two Dogmas of Empiricism. In: *The Philosophical Review* 60 (1951), S. 20–43

[Ravin 1990] RAVIN, Y.: Disambiguating and interpreting verb definitions. In: *Proceedings of the ACL 1990*. Morristown, NJ, USA: Association for Computational Linguistics, 1990, S. 260–267

[Riloff und Jones 1999] RILOFF, E.; JONES, R.: Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In: *Proceedings of AAAI/IAAI 1999*, 1999, S. 474–479

[Rissland und Daniels 1995] RISSLAND, E. L.; DANIELS, J. J.: A hybrid CBR-IR approach to legal information retrieval. In: *Proceedings of the ICAIL 1995*. College Park, Maryland, United States: ACM, 1995, S. 52–61

[Robinson 1964] ROBINSON, R.: *Definition*. Oxford, 1964

[Rosch u. a. 1976] ROSCH, E.; MERVIS, C. B.; GRAY, W. D.; JOHNSON, D. M.; BRAEM, B. P.: Basic Objects in Natural Categories. In: *Cognitive Psychology* 8 (1976), S. 382–439

[Ruiter 1993] RUITER, D.: *Institutional Legal facts. Legal Powers and Their Effects*. Dordrecht: Kluwer, 1993

[Russell u. a. 1993] RUSSELL, G.; BALLIM, A.; CARROLL, J.; WARWICK-ARMSTRONG, S.: A practical approach to multiple default inheritance for unification-based lexicons. (1993), S. 137–147

[Saggion 2004] SAGGION, H.: Identifying Definitions in Text Collections for Question Answering. In: *Proceedings of the LREC 2004*, 2004

[Saias und Quaresma 2003a] SAIAS, J.; QUARESMA, P.: Using NLP techniques to create legal ontologies in a logic programming based web information retrieval system. In: *Proceedings of the Workshop on Legal Ontologies & Web Based Legal Information Managment at ICAIL 2003*, 2003

[Saias und Quaresma 2003b] SAIAS, J.; QUARESMA, P.: A Methodology to Create Legal Ontologies in a Logic Programming Information Retrieval System. In: (Benjamins u. a., 2005), S. 185–200

[Salton 1968] SALTON, G.: *Automatic Information Organization and Retrieval*. McGraw Hill Text, 1968

[Salton und Buckley 1988] SALTON, G.; BUCKLEY, C.: Term-weighting approaches in automatic text retrieval. In: *Information Processing & Management* 24 (1988), S. 513–523

[Salton u. a. 1975] SALTON, G.; WONG, A.; YANG, C. S.: A Vector Space Model for Automatic Indexing. In: *Communications of the ACM* 18 (1975), Nr. 11, S. 613–620

[Sattelmacher und Sirp 1989] SATTELMACHER, P.; SIRP, W.: *Bericht, Gutachten und Urteil. Eine Einführung in die Rechtspraxis*. 31. München: Verlag Franz Vahlen, 1989

[von Savigny 1840] SAVIGNY, F. C. von: *System des heutigen römischen Rechts*. Berlin, 1840

[Schäfer 2007] SCHÄFER, U.: *Integrating Deep and Shallow Natural Language Processing Components – Representations and Hybrid Architectures*. Saarbrücken, Germany, Faculty of Mathematics and Computer Science, Saarland University, Dissertation, 2007

[Schiller u. a. 1999] SCHILLER, A.; TEUFEL, S.; STÖCKERT, C.; THIELEN, C.: Guidelines für das Tagging deutscher Textcorpora mit STTS / Institut für maschinelle Sprachverarbeitung. Stuttgart, 1999. – Forschungsbericht

[Schmid 1994] SCHMID, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of NMLP 1994*, September 1994

[Schmidt 2001] SCHMIDT, C.: *Zur Verständlichkeit von Allgemeinen Versicherungsbedingungen. Empirische Untersuchungen zur Linearisierung der VHB 92 (Allgemeine Hausratsversicherungsbedingungen)*, Humboldt-Universität zu Berlin, Institut für deutsche Sprache und Linguistik, Diplomarbeit, 2001

[Schwacke 2003] SCHWACKE, P.: *Juristische Methodik mit Technik der Fallbearbeitung*. Stuttgart, 2003

[Schweighofer und Geist 2007] SCHWEIGHOFER, E.; GEIST, A.: Legal Query Expansion using Ontologies and Relevance Feedback. In: (Casanovas u. a., 2008), S. 149–160

[Schweighofer und Winiwarter 1995] SCHWEIGHOFER, E.; WINIWARTER, W.: KONTERM: Exploratory Data Analysis for Semi-automatic Indexation of Legal Documents. In: *Proceedings of the DEXA Workshop 1995*, 1995, S. 407–412

[Schäfer 2005] SCHÄFER, B.: Ontological commitment and the concept of "legal system" in comparative law and legal theory. In: *ARSP Beihefte* 102 (2005), S. 141–151

[Scott Piao und Ananiadou 2008] SCOTT PIAO, J. M.; ANANIADOU, S.: Clustering Related Terms with Definitions. In: *Proceedings of the LREC 2008*. Marrakech, Morocco, may 2008

[Searle 1969] SEARLE, J.: *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press., 1969

[Sergot u. a. 1986] SERGOT, M. J.; SADRI, F.; KOWALSKI, R. A.; KRIWACZEK, F.; HAMMOND, P.; CORY, H. T.: The British Nationality Act as a logic program. In: *Communications of the ACM* 29 (1986), Nr. 5, S. 370–386

[Sierra und Alarcón 2002] SIERRA, G.; ALARCÓN, R.: Identification of Recurrent Patterns to Extract Definitory Contexts. In: GELBUKH, A. F. (Hrsg.): *Proceedings of CICLing 2002* Bd. 2276, Springer, 2002, S. 436–438

[Sierra und McNaught 2000] SIERRA, G.; MCNAUGHT, J.: Extracting semantic clusters from the alignment of definitions. In: *Proceedings of the ACL 2000*. Saarbrücken, Germany: Association for Computational Linguistics, 2000, S. 795–799

[Smith 1997] SMITH, J. C.: The use of lexicons in information retrieval in legal databases. In: *Proceedings of the ICAIL 1997*. Melbourne, Australia: ACM, 1997, S. 29–38

[Smith u. a. 1995] SMITH, J. C.; GELBART, D.; MACCRIMMON, K.; ATHERTON, B.; MCCLEAN, J.; SHINEHOFT, M.; QUINTANA, L.: Artificial Intelligence and Legal Discourse: The Flexlaw Legal Text Management System. In: *Artificial Intelligence and Law* 3 (1995), Nr. 1-2, S. 55–95

[Soderland 1999] SODERLAND, S.: Learning Information Extraction Rules for Semi-Structured and Free Text. In: *Machine Learning* 34 (1999), Nr. 1-3, S. 233–272

[Spärck Jones 1972] SPÄRCK JONES, K.: A statistical interpretation of term specificity and its application in retrieval. In: *Journal of Documentation* 28 (1972), S. 11–21

[Stevenson und Greenwood 2006] STEVENSON, M.; GREENWOOD, M. A.: Comparing Information Extraction Pattern Models. In: *Proceedings of the ACL 2006 Workshop on Information Extraction Beyond the Document*. Sydney, Australia, 2006

[Storrer und Wellinghoff 2006] STORRER, A.; WELLINGHOFF, S.: Automated detection and annotation of term definitions in German text corpora. In: *Proceedings of the LREC 2006*. Genua, Italy, 2006

[Strehlow 1983] STREHLOW, R. A.: Terminology and the Well-Formed Definition. In: INTERRANTE, C. (Hrsg.); HEYMANN, F. (Hrsg.): *Standardization*

*of Technical Terminology: Principles and Practices, ASTM STP 806*. 1983, S. 15–25

[Sudo u. a. 2003] SUDO, K.; SEKINE, S.; GRISHMAN, R.: An Improved Extraction Pattern Representation Model for Automatic IE Pattern Acquisition. In: *Proceedings of the ACL 2003*, 2003, S. 224–231

[Summers 1987] SUMMERS, D. (Hrsg.): *Longman Dictionary of Contemporary English: New Edition*. Essex: Longman, 1987

[Tomlinson u. a. 2006] TOMLINSON, S.; LEWIS, D. D.; OARD, D. W.: TREC 2006 Legal Track Overview. In: *Proceedings of TREC 2006*, 2006

[Tomlinson u. a. 2007] TOMLINSON, S.; OARD, D. W.; BARON, J. R.; THOMPSON, P.: Overview of the TREC 2007 Legal Track. In: *Proceedings of TREC 2007*, 2007

[Trimble 1985] TRIMBLE, L.: *English for science and technology.* Cambridge, 1985

[Turmo u. a. 2006] TURMO, J.; AGENO, A.; CATALÀ, N.: Adaptive information extraction. In: *ACM Computing Survey* 38 (2006), Nr. 2

[Turtle 1995] TURTLE, H.: Text retrieval in the legal world. In: *Artificial Intelligence and Law* 3 (1995), March, Nr. 1, S. 5–54

[Valente 1995] VALENTE, A.: *Legal knowledge engineering, a modelling approach*. Amsterdam, University of Amsterdam, Dissertation, 1995

[Valente und Breuker 1994] VALENTE, A.; BREUKER, J.: A functional ontology of law. In: *G. Bargellini and S. Binazzi, editors, Towards a global expert system in law. CEDAM Publishers, Padua, Italy, 1994.* 1994

[Vandenberghe u. a. 2003] VANDENBERGHE, W.; SCHÄFER, B.; KINGSTON, J.: Ontology Modelling in the Legal Domain - Realism Without Revisionism. In: GRENON, P. (Hrsg.); MENZEL, C. (Hrsg.); SMITH, B. (Hrsg.): *Proceedings of the KI Workshop on Reference Ontologies and Application Ontologies 2003* Bd. 94, CEUR-WS.org, 2003

[Visser und Bench-Capon 1996] VISSER, P.; BENCH-CAPON, T.: The formal specification of a legal ontology. In: *Proceedings of JURIX 1996*, 1996

[Voorhees 2003] VOORHEES, E. M.: Overview of the TREC 2003 Question Answering Track. In: *Proceedings of TREC 2003*, 2003, S. 54–68

[Voorhees 2004] VOORHEES, E. M.: Overview of the TREC 2004 Question Answering Track. In: *Proceedings of TREC 2004*, 2004

[Vossen 1998] VOSSEN, P. (Hrsg.): *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers, 1998

[Vossen und Copestake 1993] VOSSEN, P.; COPESTAKE, A.: Untangling definition structure into knowledge representation. In: (Briscoe u. a., 1993), S. 246–274

[Wagner 1972] WAGNER, H.: *Die Deutsche Verwaltungssprache der Gegenwart. Eine Untersuchung der sprachlichen Sonderform und ihrer Leistung*. Düsseldorf, 1972

[Waismann 1945] WAISMANN, F.: Verifiability. In: *Proceedings of the Aristotelean Society* 19 (1945), S. 119–50

[Walker 1998] WALKER, R.: *Die Publikation von Gerichtsentscheidungen*, Universität des Saarlandes, Dissertation, 1998

[Walter 2006] WALTER, S.: *Richtlinien für die Annotation von Definitionen im Projekt CORTE*. Web-Dokument, `http://www.coli.uni-saarland.de/projects/corte/walter06richtlinien.pdf`, 18. Juni 2010. 2006

[Walter 2009] WALTER, S.: Definition Extraction from Court Decisions Using Computational Linguistic Technology. In: GREWENDORF, G. (Hrsg.); RATHERT, M. (Hrsg.): *Formal Linguistics and Law*. Berlin: Mouton de Gruyter, 2009

[Weinberger 1979] WEINBERGER, O.: Kann man das normenlogische Folgerungssystem philosophisch begründen? In: *ASRP* (1979), S. 178

[Westerhout und Monachesi 2007] WESTERHOUT, E.; MONACHESI, P.: Combining pattern-based and machine learning methods to detect definitions for eLearning purposes. In: *Proceedings of the workshop on Natural Language Processing and Knowledge Representation for eLearning Environments at RANLP 2007*. Sofia, Bulgaria, 2007

[Wilks u. a. 1995] WILKS, Y. A.; SLATOR, B. M.; GUTHRIE, L. M.: *Electric words : dictionaries, computers, and meanings*. Cambridge, Mass. [u.a.]: MIT Press, 1995

[Xu 2007] XU, F.: *Bootstrapping Relation Extraction from Semantic Seeds*, Saarland University, Dissertation, 2007

[Xu u. a. 2003] XU, J.; LICUANAN, A.; WEISCHEDEL, R. M.: TREC 2003 QA at BBN: Answering Definitional Questions. In: *Proceedings of TREC 2003*, 2003, S. 98–106

[Xu u. a. 2004] XU, J.; WEISCHEDEL, R. M.; LICUANAN, A.: Evaluation of an extraction-based approach to answering definitional questions. In: SANDERSON, M. (Hrsg.); JÄRVELIN, K. (Hrsg.); ALLAN, J. (Hrsg.); BRUZA, P. (Hrsg.): *Proceedings of SIGIR 2004*, ACM, 2004, S. 418–424

[Xu u. a. 2005] XU, J.; CAO, Y.; LI, H.; ZHAO, M.: Ranking definitions with supervised learning methods. In: ELLIS, A. (Hrsg.); HAGINO, T. (Hrsg.): *Proceedings of WWW 2005 (Special interest tracks and posters)*, ACM, 2005, S. 811–819

[Xu u. a. 2006] XU, J.; CAO, Y.; LI, H.; ZHAO, M.; HUANG, Y.: A Supervised Learning Approach to Search of Definitions. In: *Journal of Computer Science and Technology* 21 (2006), Nr. 3, S. 439–449

[Yang u. a. 2003] YANG, H.; CUI, H.; MASLENNIKOV, M.; QIU, L.; KAN, M.-Y.; CHUA, T.-S.: QUALIFIER In TREC-12 QA Main Task. In: *Proceedings of TREC 2003*, 2003, S. 480–488

[Yeh 2000] YEH, A. S.: More accurate tests for the statistical significance of result differences. In: *Proceedings of COLING 2000*, Morgan Kaufmann, 2000, S. 947–953

[Yu u. a. 2007] YU, H.; LEE, M.; KAUFMAN, D.; ELY, J.; OSHEROFF, J. A.; HRIPCSAK, G.; CIMINO, J. J.: Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians. In: *Journal of Biomedical Informatics* 40 (2007), Nr. 3, S. 236–251

[Zadeh 1965] ZADEH, L. A.: Fuzzy Sets. In: *Information and Control* 8 (1965), Nr. 3, S. 338–353

[Zadeh 1974] ZADEH, L. A.: Fuzzy Logic and Its Application to Approximate Reasoning. In: *Proceedings of the IFIP Congress 1974*, 1974, S. 591–594

[Zhang u. a. 2005] ZHANG, Z.; ZHOU, Y.; HUANG, X.; WU, L.: Answering Definition Questions Using Web Knowledge Bases. In: DALE, R. (Hrsg.); WONG, K.-F. (Hrsg.); SU, J. (Hrsg.); KWONG, O. Y. (Hrsg.): *Proceedings of the IJCNLP 2005* Bd. 3651. Jeju Island, Korea: Springer, 2005, S. 498–506

[Zippelius 2005] ZIPPELIUS, R.: *Juristische Methodenlehre*. München, 2005